Advice for the Steady:
Decision Theory and the Requirements of Instrumental Rationality

by

Johanna Marie Thoma

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Philosophy
University of Toronto

Abstract

Advice for the Steady:
Decision Theory and the Requirements of Instrumental Rationality

Johanna Marie Thoma
Doctor of Philosophy
Graduate Department of Philosophy
University of Toronto
2017

Standard decision theory, or rational choice theory, is often interpreted to be a theory of instrumental rationality. This dissertation argues, however, that the core requirements of orthodox decision theory cannot be defended as general requirements of instrumental rationality. Instead, I argue that these requirements can only be instrumentally justified to agents who have a desire to have choice dispositions that are stable over time and across different choice contexts. Past attempts at making instrumentalist arguments for the core requirements of decision theory fail due to a pervasive assumption in decision theory, namely the assumption that the agent's preferences over the objects of choice – be it outcomes or uncertain prospects – form the standard of instrumental rationality against which the agent's actions are evaluated. I argue that we should instead take more basic desires to be the standard of instrumental rationality. But unless agents have a desire to have stable choice dispositions, according to this standard, instrumental rationality turns out to be more permissive than orthodox decision theory.

Acknowledgements

I owe an immeasurable debt of gratitude to Sergio Tenenbaum. He has been the best supervisor any student of philosophy could hope for. Every part of this dissertation has improved immensely due to his thorough, piercing, and always helpful feedback. Moreover, he inspired much of my thinking in these pages, some of which goes back to a graduate seminar on diachronic rationality that I took with him and Julia Nefsky in 2013. No less important, he always made philosophy fun.

Jonathan Weisberg, Joe Heath and Julia Nefsky also read drafts of all parts of this dissertation, and helped me improve it greatly both through thorough written feedback, and in many fruitful and inspiring discussions. Since I was on a tight schedule, I am especially grateful to my entire committee for the speed at which they read and commented on earlier drafts. They also helped me explore a wide range of interesting literature in preparation for writing this dissertation. I thank Jonathan Weisberg in particular, since work on a joint paper with him on risk-weighted expected utility theory was also extremely fruitful for my dissertation work, and so was writing a handbook article on decision theory for the *Open Handbook of Formal Epistemology* that he is editing with Richard Pettigrew. Parts of Chapters 2 and 5 contain passages from that handbook article.

This dissertation also benefitted crucially from a research stay at Stanford University in early 2016. I am extremely grateful to the Balzan Foundation for funding this visit, and to Michael Bratman for agreeing to host me, and reading and discussing my work with me extensively during that time. Several of the core ideas in this dissertation were developed while I was taking his seminar on plan rationality. I also thank the attendants of that seminar, in particular John Broome, with whom I discussed some of these ideas.

Parts of this dissertation were presented at various conferences and departmental colloquia. I thank audiences at the European Congress in Analytical Philosophy in Bucharest, the University of Konstanz, the Canadian Philosophical Association Meeting in Ottawa, the Joint Sessions of the Aristotelian Society and Mind Association in Warwick, the Workshop on the Ethics of Social Risk in Montreal, the Varieties of Agency Workshop at Stanford University, the Meeting of the Society for Exact Philosophy in Miami, the Saint Louis Annual Conference on Reasons and Rationality, the Cambridge Philosophy of Science Seminar, the University of Bristol, the London Judgment and Decision Making Group at University College London, the Choice Group at the London School of Economics, the University of Kent, the Moral Philosophy Seminar at the University of Oxford, and the Erasmus Institute for Philosophy of Economics at Erasmus University Rotterdam for helpful feedback and suggestions. Discussions with Luc Bovens, Richard Bradley, and Bernhard Salow proved especially useful.

The philosophy community at the University of Toronto provided a wonderful environment in which to develop and abandon philosophical ideas. My friends and fellow graduate students always challenged and inspired me. They were also an invaluable source of support and distraction. I miss them dearly. I also thank the Connaught Fund for sponsoring my doctoral research at the University of Toronto, and the administrators at the Philosophy Department for their always cheerful and efficient support.

By hiring me, the Department of Philosophy, Logic, and Scientific Method at the London School of Economics forced me to write this dissertation quickly. For that, and for obvious other reasons, I am extremely grateful to them.

Finally, I wish to thank my parents, for everything, and Marius Backmann, for putting up with me at all stages of writing this dissertation, and for lifting my mind from it whenever I needed it most.

Contents

1 Introduction 1

2 Maximization and Temptation 5

2.1 Introduction 5

2.2 Standard Requirements of Decision Theory 6

2.3 Preference-Based Instrumental Rationality 9

2.4 A Temptation Problem 10

2.5 Instrumentalist Two-Tier Arguments 12

2.6 Problems with Instrumentalist Two-Tier Arguments 16

2.7 Time-Slice Cooperation Arguments 18

2.8 Problems with Time-Slice Cooperation Arguments 20

2.9 Giving up Preference-Based Instrumental Rationality 21

2.10 Conclusions 23

3 Preference-Based Instrumental Rationality and the Money Pump Argument 25

3.1 Introduction 25

3.2 Cyclical Preferences 26

3.3 The Money Pump Argument 27

3.4 Preference-Based Instrumental Rationality and the Money Pump Argument 28

3.5 The True Cost of Being Money Pumped 33

3.6 Desire-Based Instrumental Rationality 34

3.7 Desire-based Instrumental Rationality and Instrumental Arguments 38

3.8 Conclusions 43

4 Desire-Based Instrumental Rationality and the Money Pump Argument 44

4.1 Introduction 44

4.2 Foresight as an Alternative Response to Money Pumps 45

4.3 Preference as a Summary of Desire 50

4.4 Uniqueness and Preference-Guidance 51

4.5 Back to the Money Pump Argument 53

4.6 Uniqueness and Maximization 56

4.7 Preference as Disposition to Choose 60

4.8 Money Pumps and the Stability of Preference 63

4.9 The Desire for Stability 66

4.10 Conclusions 69

5 Decision under Uncertainty 71

5.1 Introduction 71

5.2 Expected Utility Theory 72

5.3 The Allais Paradox 76

5.4 The Dynamic Allais Problem 78

5.5 Consequentialism 80

5.6 Prospects and the Standard of Instrumental Rationality 83

5.7 Dynamic Dominance Violations 89

5.8 Resolution 92

5.9 Conclusions 94

6 Risk Aversion and the Long Run 96

6.1 Introduction 96

6.2 Samuelson's Colleague 97

6.3 Risk-Weighted Expected Utility Theory 100

6.4 Risky Choice over Time 104

6.5 Framing Decision Problems 108

6.6 Risk-weighted Expected Utility Theory and Framing 110

6.7 Permissiveness About Dynamic Choice 113

6.8 Conclusions 117

7 Concluding Remarks 119

Bibliography 121

List of Tables

4.1 Partial Rankings of Outcomes 58

5.1 State-outcome matrix 72
5.2 Should I cycle to work? 72
5.3 Allais Paradox: First Choice 77
5.4 Allais Paradox: Second Choice 77
5.5 Allais Paradox: Second Choice 88

List of Figures

2.1 Temptation Problem . . . 11
2.2 Inter-Temporal Prisoner's Dilemma . . . 14

4.1 Dynamic Money Pump Problem . . . 46
4.2 Modified Dynamic Money Pump Problem . . . 47

5.1 Dynamic Allais Problem . . . 79
5.2 Alternative Second Choice . . . 90

6.1 A Concave Utility Function . . . 99
6.2 Certainty Equivalent Per Gamble for Repetitions of SC's Gamble . . . 103
6.3 Dynamic Version of SC's Decision Problem, Two Consecutive Choices . . . 105

Chapter 1

Introduction

Instrumental rationality requires agents to take the best means to the ends they desire. Most decision theorists assume that standard normative decision theory is concerned with instrumental rationality, and with instrumental rationality alone. For instance, Joyce (1999) claims the following in the opening paragraph of his *The Foundations of Causal Decision Theory*:

> The overarching goal of normative decision theory is to establish a general standard of rationality for the sort of *instrumental* (or "practical") reasoning that people employ when trying to choose means appropriate for achieving ends they desire. (p.9)

Buchak (2014) calls expected utility theory, standard decision theory in the context of uncertainty, the "orthodox theory of instrumental rationality" (p.1091). For the most part, the assumption that decision theory is a theory of instrumental rationality is so entrenched, it is rarely stated or explicitly argued for.[1]

The position that instrumental rationality is all there is to practical rationality is often described as a Humean notion of rationality, following Williams (1979). In the *Treatise of Human Nature*, Hume famously asserted that "reason is, and ought only to be the slave of the passions [...] Where a passion is neither founded on false supposition, nor chooses means insufficient for the end, the understanding can neither justify nor condemn it." (Hume (2007/1739), II.3.3 415-416) According to what is now known as the Humean theory of practical rationality, while agents should take the best means to their ends, rationality is silent on what ends agents ought to hold.[2]

Those who think that normative decision theory is (only) about instrumental rationality need not subscribe to such Humeanism. The idea is merely that, if there are non-instrumental requirements of rationality, those concern what ends an agent should have, and that such concerns are outside the realm of decision theory. Decision theory is in that sense Humean, but decision theory need not be all there is to practical reason.

This dissertation is concerned with whether the standard requirements of orthodox decision theory can be interpreted as requirements of instrumental rationality. In particular, the requirements I will

[1] For examples of papers whose premise is based on this assumption, see Verbeek (2001) or Gaus (2008). Lewis (1988) defends Humeanism about decision theory by arguing that one major form of non-Humeanism is not compatible with decision theory — but he never positively argues that decision theory is a Humean theory of rationality.

[2] While this instrumental notion of practical rationality has been popular, it is not uncontroversial that Hume himself actually held it. See Hampton (1995) for arguments that he did not. Also note that a similar notion of instrumental rationality can already be found in Hobbes' *Leviathan*, Hobbes (2010/1651).

consider are the requirements to maximize with regard to one's preferences (see Section 2.2), the requirement to have well-ordered, that is, transitive and complete, preferences (see also Section 2.2), and the requirement to have separable preferences in the context of uncertainty, that is, follow a version of Savage's (1954) sure-thing principle (see Section 5.2). The latter is characteristic of expected utility theory, while the former two are presupposed by most formal decision theories.

I will argue that these core requirements of orthodox decision theory cannot be defended as general requirements of instrumental rationality. Instead, I show that these requirements can only be justified as conditional requirements of instrumental rationality: They turn out to be requirements of instrumental rationality for agents who have a desire to have choice dispositions that are stable over time and across different choice contexts.

Some requirements of standard decision theory — such as the requirement that agents should maximize, or that agents should be sophisticated in dynamic choice contexts (see Section 4.2) — have appeared to many to be obvious requirements of instrumental rationality. Others, such as transitivity or separability, have been defended as requirements of instrumental rationality by appealing to various instrumentalist arguments. What these arguments typically have in common is that they point out that agents who violate those requirements are prone to making a sure loss in some choice scenarios. Those arguments, as I will show, typically take the intuitive plausibility of maximization and sophistication for granted.

This dissertation aims to establish that at the heart of this joint instrumentalist defence of the various requirements of orthodox decision theory lies an equivocation about the standard of instrumental rationality. It is a pervasive assumption in decision theory that the agent's preferences over the objects of choice — be it outcomes or uncertain prospects — form the standard of instrumental rationality against which the agent's actions are evaluated. The requirement that our choices ought to be guided by our preferences, as captured, for instance, by the standard requirement to maximize, is plausible according to this preference-based notion of instrumental rationality. However, I will argue that the instrumentalist arguments for further requirements, such as transitivity or separability, fail according to this standard. Their appeal in fact relies on a different understanding of the standard of instrumental rationality, one that relates to more basic desires. But according to this standard, the requirement to maximize is no longer justifiable.

The equivocation about the standards of instrumental rationality thus seriously calls into question the instrumentalist defence of standard decision theory. In light of this, I develop what I take to be the best case that can be made for the standard requirements of decision theory. I argue that, for independent reasons, it is more plausible to take more basic desires rather than preferences over the objects of choice to be the standard of instrumental rationality. According to this standard, we can show that instrumental rationality may require agents who have a desire to have stable choice dispositions to abide by the core requirements of orthodox decision theory. For all other agents, however, instrumental rationality turns out to be more permissive than orthodox decision theory.

Chapters 2, 3 and 4 deal with decision-making in the context of certainty, when an agent knows what the outcomes of the actions open to her will be. In this context, standard decision theory requires agents to have transitive and complete preferences over those outcomes, and to maximize with regard to them. Chapter 2 is concerned with the maximization requirement. This requirement has been called into question by authors who would like to argue that it can be rational to resist temptation, where temptations are understood as temporary shifts in our preferences. When an agent makes several choices in a context where she is subject to temptation, she appears to do better by her own lights if she adopts a choice strategy that allows her to act counter-preferentially and resist the temptation. It is hence argued that in temptation cases, instrumental rationality demands that agents violate the maximization requirement. I show that this argument in favour of resisting temptation fails under the assumption that preferences over outcomes form the standard of instrumental rationality, which its proponents are committed to. But if we give up the assumption, the arguments are redundant, save for a special case. And that is because, if preferences are not themselves the standard of instrumental rationality, they can misrepresent the true standard of instrumental rationality. In that case, resisting temptation and failing to maximize with regard to one's tempted preferences can be rational for straightforward reasons. Maximization can then only be a principle of rationality that is conditional on preferences accurately reflecting the true standard of instrumental rationality.

Chapters 3 and 4 are concerned with a famous instrumentalist argument in favour of the acyclicity of preference, which is strictly weaker than transitivity. And that is the money pump argument, which holds that any agents with cyclical preferences can be engaged in a series of trades that leaves her with what she started with having lost something she desires. To avoid this, the agent should adopt acyclical preferences. Chapter 3 establishes that, while this argument is usually fleshed out appealing to preferences over outcomes as the standard of instrumental rationality, the argument must fail according to that standard. Instead, the appeal of those arguments implicitly relies on a notion of instrumental rationality according to which more basic desires for simpler states of affairs which are features of full outcomes form the standard of instrumental rationality. I defend this alternative notion of instrumental rationality and argue that it can explain what is instrumentally irrational about being money pumped.

Chapter 4 explores whether the money pump argument can justify a requirement to have acyclical preferences given this alternative, desire-based notion of instrumental rationality. While we have established that being money pumped is (often) instrumentally irrational, the challenge to this argument is that there may be alternative ways of avoiding being money pumped other than adopting acyclical preferences. In particular, agents may fail to maximize with regard to their preferences at crucial points in time, or adjust their preferences temporarily to avoid being money pumped, while keeping their cyclical preferences for the most part. I argue that desire-based instrumental rationality allows for these alternative responses since it struggles to justify the requirement of maximization, and appears to allow for multiple preference relations to be equally permissible given the agent's desires. The best case that can be made for acyclicity involves reinterpreting preferences as dispositions to choose. Acyclicity can then be justified to agents who have a desire to have dispositions to choose that are stable over time and across different choice contexts. And this is because the alternative responses to the money pump argument would frustrate that desire.

Chapters 5 and 6 turn to choice under uncertainty. The central requirement of orthodox decision theory under uncertainty, or expected utility theory, is the requirement of separability. Roughly, the idea behind separability is that an agent's preferences over two prospects should not be affected by what happens in states of the world that are not part of those prospects. Chapter 5 shows that for reasons that are parallel to those that applied in the case of acyclicity, this requirement, too, can only be defended as a conditional requirement of instrumental rationality. The standard instrumentalist argument that is made in favour of separability appeals to a dynamic choice context in which agents who violate separability stand to make a sure loss. I first show that this argument, like the money pump argument, fails under the common assumption that preferences over the objects of choice — in this case uncertain prospects —

form the standard of instrumental rationality. In fact, for us to establish even the most uncontroversial principles of choice in the context of uncertainty, we must take the standard of instrumental rationality to have only outcomes, rather than prospects directly, as its object. If that is so, however, instrumental rationality again turns out more permissive than orthodox decision theory implies. In particular, agents can again avoid the instrumental irrationality of sure loss without adopting separable preferences, by either acting counter-preferentially, or by making temporary adjustments to their preferences.

Finally, Chapter 6 considers a recent alternative to expected utility theory that relaxes the separability requirement and is more permissive about choice under uncertainty. This theory, Buchak's (2013) risk-weighted expected utility theory, thus holds the promise of both capturing the fact that instrumental rationality is more permissive than standard expected utility theory implies, as well as still providing a formalism that is supposed to capture, explain and predict how different agents tend to behave in the face of risk. This theory is plausible when we focus on one-off decisions. However, I argue that, once we take into account the fact that agents face many risky choices in their lives, this theory either ends up making similar recommendations as expected utility theory, or ends up being extremely sensitive to the way in which agents specify their decision problems, and choose to act in dynamic choice contexts. This calls into question the usefulness of the formalism. Similar claims should apply for any theory that tries to account for common types of non-separable preferences.

The dissertation thus concludes that the core requirements of orthodox decision theory can be justified instrumentally to agents who have a desire to have choice dispositions that are stable over time and across different choice contexts. But for all others, instrumental rationality turns out to be more permissive. Moreover, in the face of this permissiveness, the usefulness of any formalism to capture agents' choices in the context of uncertainty is called into question.

We started by noting that most decision theorists are Humeans about decision theory. There is good reason to care whether the core requirements of orthodox decision theory can be given an instrumentalist justification, even aside from this contingent fact. And that is that, if we could give such a justification, we would have a response to the charge that these requirements express nothing but a fetish for consistency or psychic tidiness.[3] Suppose somebody asks, "Why should I be consistent in the way your decision theory says I should be?" If we could provide instrumentalist justifications for the core requirements of the decision theory, we could respond, "Because this is the best way to serve your ends."

The arguments presented in this dissertation provide us only with a conditional instrumental defence of the core requirements of decision theory. The response to those who wonder about why they should be consistent in the sense that decision theory requires them to be is correspondingly unsatisfying. All we can say to them, I argue, is that this is one way of serving their ends well. But unless they have the desire to have stable choice dispositions, there will be other ways. There is thus a sense in which agents already need to be concerned with consistency, namely consistency of choice over time and across choice contexts, in order to have conclusive reason to abide by further consistency requirements of orthodox decision theory.

[3] Kolodny (2005, 2007) poses this challenge to requirements of rationality more generally.

Chapter 2

Maximization and Temptation

2.1 Introduction

When we are certain about the consequences of our actions, standard decision theory requires us to maximize with regard to our preferences over the outcomes of the actions available to us. Most authors writing on decision theory also assume what I will call *preference-based instrumental rationality.* That is, they assume that preferences are the fundamental conative attitude against which actions are assessed instrumentally. They are the standard of instrumental rationality. If we assume *preference-based instrumental rationality*, *maximization* appears to be a straightforward requirement of instrumental rationality.

The requirement to maximize has been called into question as a requirement of instrumental rationality, however. In particular, it has been attacked by authors who would like to argue that it can be rational to resist temptation, where temptations are understood as temporary shifts in our preferences. We all seem to be prone to such temporary shifts in our preferences. I might plan, for instance, to only stream one episode of a TV show during my coffee break. But then, once I have watched the first, I come to prefer to watch another. Temptation cases, in the following, are understood to be cases where an agent's preferences temporarily shift and then return to what they were before.

Standard decision theory does not usually condemn such changing preferences as irrational per se. Indeed, this seems to be as it should be, if decision theory is to express a Humean notion of rationality. Hume claimed that "[‘t]is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger." (Hume (2007/1739), II.3.3 416). So an agent who prefers the destruction of the world to the scratching of her finger is equally rational as a person who prefers the scratching of her finger to the destruction of the world. Then why should it be irrational if the same agent had one preference on one day, and another preference on a different day?[1]

Temptation cases are said to challenge *maximization* in the following way. When an agent makes

[1] For a defence of the idea that there is no relevant difference between attitudes of the same agent at different points in time and attitudes of different agents, as far as rationality is concerned, see Hedden (2015a). He defends what he calls *time-slice rationality*, the claim that all requirements of rationality are requirements on an agent's attitudes and choices at a particular point in time. I am not committed to this strong claim. My point here is merely that changing preferences are not irrational per se. This keeps open the possibility that they are sometimes irrational because of costs they may bring for the agent. This possibility will be considered in the next chapters. Note that it also keeps open the possibility that changing preferences are sometimes rational because of the benefits they bring to the agent. Indeed, several authors have pointed out that inconstancy of preference may sometimes be the key to a successful life. See, for instance, Blackburn (1995) and Bovens (1999).

several choices in a context where she is subject to temptation, she appears to do better by her own lights if she adopts a choice strategy that allows her to act counter-preferentially. In particular, in cases of temptation, she sometimes does better by making a resolution to resist the temptation — e.g. to only watch one episode — and then go through with the resolution, against the temporary preference to give into the temptation. Since this kind of choice strategy has instrumental advantages, the agent should adopt it. And so it is argued that in temptation cases, instrumental rationality in fact demands that agents violate the requirement that they ought to choose an act that leads to one of their most preferred outcomes.

In this chapter, I want to argue that the two prominent ways of making this argument in favour of resisting temptation fail under the assumption of *preference-based instrumental rationality*, which the proponents of the argument are committed to. But if we give up *preference-based instrumental rationality*, the argument is redundant, save for a special case. And that is because, if preferences are not themselves the standard of instrumental rationality, they can misrepresent the true standard of instrumental rationality. In that case, resisting temptation and failing to maximize with regard to one's tempted preferences can be rational for straightforward reasons. Maximization can then at best be a principle of rationality that is conditional on preferences accurately reflecting the true standard of instrumental rationality.

Before I can make this argument, I will explain what I take to be the central requirements of orthodox decision theory in the context of certainty, along with the preference-based notion of instrumental rationality typically invoked to justify them.

2.2 Standard Requirements of Decision Theory

Standard decision theory is normally understood to require an agent to 'maximize utility' in the context of certainty. In the context of certainty, each possible action an agent might take is associated with one outcome. Outcomes, in turn, are usually taken to be complete descriptions of anything the agent may care about in the circumstances the action brings about. To maximize utility, the agent chooses the act, or one of the acts, that leads to an outcome with highest utility.

In modern decision theory, utility has come to be understood to be a mere representation of an agent's binary preferences.[2] In fact, central to modern decision theory are representation theorems that show that, if the agent's preferences abide by certain axioms, then they can be represented by a utility function that we then require the agent to maximize. This tradition goes back at least to Ramsey (1928/1950), and became mainstream in economics with von Neumann and Morgenstern (1944).

In the context of certainty, if the agent's binary preferences over outcomes form a *weak ordering* of the set of outcomes, then they can be represented by a utility function.[3] In fact, there will be a family

[2] I will, for now, take the concept of a preference to be primitive. In order for *preference-based instrumental rationality* to make sense, we must take it to be a kind of conative attitude, or pro-attitude, that could serve as the standard of instrumental rationality. In Chapter 4, I will discuss in more detail what we could mean by 'preference' and consider, in particular, the possibility that preferences might be dispositions to choose rather than conative attitudes (see Section 4.7).

[3] Depending on one's favourite decision theory, the agent may be required to have preferences that form a weak ordering over more than full outcomes. As we will see, in the context of uncertainty, where each action may lead to several different outcomes, the agent is also usually required to have preferences that form a weak ordering over uncertain prospects or lotteries. Jeffrey's (1965/1983) decision theory requires the agent to have preferences not only over outcomes as we understand them in the following, but over any proposition that could be true or false about the world. Here, we consider only *weak ordering* over outcomes. This is because that is the crucial condition when it comes to choice under certainty, where the objects of choice are in fact outcomes. Moreover, the instrumental argument in favour of transitivity we will discuss in Chapters 3 and 4 works only as an argument in favour of the transitivity of the possible objects of choice.

of utility functions that represent them. Let X be the set of outcomes the agent's actions may bring about. Let $\succcurlyeq$ represent weak preference between outcomes: $x \succcurlyeq y$ if and only if the agent either strictly prefers x to y, or is indifferent between x and y. Now $\succcurlyeq$ forms a weak ordering over X if the following two conditions are met:

Transitivity: For all x, y and $z \in X, x \succcurlyeq y$ and $y \succcurlyeq z$ implies that $x \succcurlyeq z$

Completeness: For all x and $y \in X, x \succcurlyeq y$ or $y \succcurlyeq x$

Given these two conditions, there is a utility function $U(X)$ such that an agent weakly prefers one outcome over another just in case it has at least as high a utility:

$$\text{For all } x \text{ and } y \in X,\, U(x) \geq U(y) \text{ if and only if } x \succcurlyeq y$$

This result is fairly intuitive. Transitivity and Completeness imply that the agent can form a preference ranking of all the outcomes in X, where an outcome higher up in the ranking will be preferred to all the outcomes that are lower in the ranking (though outcomes can have equal rank when the agent is indifferent between them). Given such a ranking, we can simply assign higher numbers to the outcomes that are ranked more highly. This assignment of numbers can then be used to construct a utility function that represents the agent's preferences.

We said that the central requirement of standard decision theory is usually taken to be that agents should 'maximize utility'. If utility is just a representation of binary preference in the way we just sketched, and the agent's preferences form a weak ordering of outcomes, the requirement to 'maximize utility' is simply the requirement to choose an action that leads to an outcome that is ranked most highly in the agent's preference ordering. The following requirement ensures that an agent with preferences that form a weak ordering chooses a most highly ranked outcome, or an outcome with highest utility:

Maximization: Agents ought to choose an action with an associated outcome such that no other available action leads to an outcome that is strictly preferred to it.

While standard decision theory under certainty is often summed up under the slogan 'maximize utility', maximization is not the only requirement that is taken to be a requirement of rationality in normative decision theory. In particular, some or all of the axioms of the representation theorems are also often taken to be requirements of rationality. In the context of certainty, it is usually taken to be a requirement of rationality that an agent's preferences ought to be transitive, and less often that they ought to be complete.[4] That is, it is required that the agent's preferences form a weak ordering over outcomes. Let me call this requirement *weak ordering*:

Weak Ordering: Agents ought to have preferences that form a weak ordering over outcomes.

Given we understand utility simply as a representation of binary preference, *weak ordering* and *maximization* are the central requirements of standard normative decision theory in the context of certainty. But utility is sometimes also understood in a 'realist' sense, that is, as more than a mere representation of preference. In fact, this was the norm before the development of the representation theorems just mentioned. Classical utilitarians such as Bentham (1789/2007) and Mill (1861/1998), for

[4] Several authors have argued that completeness is too strong a requirement when the outcome space is large. Instead, it is enough if an agent's preferences are coherently extendible, that is, if there is a way of completing the agent's preferences that would conform with the other axioms. See, for instance, Joyce (1999).

instance, thought of utility as pleasure and the absence of pain. However, such a definition of utility would not help us cast decision theory as a theory of instrumental rationality. If I were to be required to maximize utility thus understood, I would be rationally required to maximize my own pleasure, even if I did not care about pleasure and had no desire to maximize it. The same applies to definitions of utility as wellbeing understood more broadly.[5]

For a realist interpretation of utility to serve as part of an instrumental theory of rationality, utility must be a measure of the agent's desires or preferences. We could perhaps understand utility as something like 'desiredness', where different quantities of desiredness express the varying degrees to which an agent desires outcomes. Jeffrey (1965/1983), for instance, speaks of desirabilities instead of utilities, and interprets them as degrees of desire (p.63). We might think that once we understand utility in realist terms, the only requirement of rationality in the context of certainty is *maximization*. If we have independent access to utility, we are not in need of representation theorems. Moreover, orderings according to some quantity are necessarily transitive. Therefore, preferences would end up being transitive as long as they track utility.

Indeed, it is sometimes claimed that we can justify the transitivity of preference because preferences should rationally be required to track some property that is transitive as a matter of logic. For instance, the betterness relation is usually taken to be transitive as a matter of logic.[6] If preference was rationally required to track betterness, then this would give us an argument that preferences should be transitive. Similarly, Hedden (2015a) makes an argument inspired by Kolodny (2008) to the effect that preferences ought to track the 'having more reason to desire a rather than b' relation, and that this relation is transitive.[7]

The only way in which we could cast such an argument in terms of instrumental rationality is if the property that preferences are required to track is itself an expression of the agent's desires, such as 'being more desired than'. Can we make such an argument by appealing to a realist understanding of utility as 'desiredness'? If desiredness was a quantity, where having more or less of this quantity corresponds to different degrees of desiredness, then we could make such an argument. But I take this not to be obvious. In fact, the centrality of representation theorems in decision theory serves as some evidence that this is not obvious.[8] As we will see in the following, if we are worried about the transitivity of preference, we are worried precisely about whether an agent can order all outcomes in a single ranking in terms of her desires. Such worries would directly translate into worries about whether the agent's desires can be captured by a single quantity of 'desiredness'. And so an argument for transitivity that appeals to such a quantity would be question-begging.

That transitivity holds is thus something that needs to be argued for before we can defend a realist

[5] Note, in particular, that a theory of instrumental rationality is different from what Parfit (1984) calls the Self-Interest Theory of Rationality. According to instrumental rationality, the agent's actions are evaluated in terms of her desires or preferences, not in terms of whether they serve her self-interest or wellbeing — unless she desires her own wellbeing.

[6] See, for instance, Broome (1991). For an opposing view, see Temkin (2012).

[7] In response to Portmore (2011), who bases his consequentialism on the 'more reason to desire' relation, Tenenbaum (2014b) argues that this relation cannot in fact explain our reasons for action: For instance, I may have more reason to desire the welfare of my child than I have reason to desire to give a job in my department to the most qualified candidate, while still having reason not to give the job to my child when there is a more qualified candidate, but my child would greatly benefit from getting it. More plausibly, my reasons for action should track what I have reason to desire more. But what I have reason to desire more cannot explain the transitivity of comparative desire, or preference.

[8] The centrality of the representation theorems is arguably also the historical baggage of a more general positivist and behaviourist turn in the social sciences in the early and mid-20th century. See, for instance, Witt (2005) for a historical account. Strict positivism is largely abandoned now, and even within economics there are some signs of a return to a hedonic notion of utility. See, for instance, Kahnemann et al. (1997). However, for the reasons just given, this return would also seem to be a return to a self-interest understanding of rationality. If decision theory is supposed to capture instrumental rationality, then utility needs to be a measure of preference or desire.

interpretation of utility. Either way, I thus take *weak ordering* and *maximization* to be the central requirements of standard normative decision theory in the context of certainty. If normative decision theory is supposed to capture the requirements of instrumental rationality, and only those, then we need to show that both of these requirements are requirements of instrumental rationality.

In the following, I will consider both the standard reasons for accepting *maximization*, and the most common reason for rejecting it. For now, I assume that the agent's preferences abide by *weak ordering.* The next two chapters will consider instrumentalist arguments for the transitivity requirement of *weak ordering.* Notably, justifications as well as criticism of *maximization* rely on what I will call, in the following, *preference-based instrumental rationality.* The next chapter will show that, if we want to have any hope of justifying transitivity instrumentally, we have to give up that assumption.

2.3 Preference-Based Instrumental Rationality

Instrumental rationality is traditionally understood as requiring agents to take the best means to ends they desire. But note that ends and desires do not appear in decision theory as we depicted it. Instead, the theory features binary preferences. In the context of certainty, where agents are choosing between outcomes, the relevant preferences are preferences over these outcomes. So how could the requirements of standard decision theory, like *maximization*, be understood as requirements of instrumental rationality?

On a broader understanding of instrumental rationality, actions or principles of choice are evaluated in light of the agent's own conative attitudes, or pro-attitudes.[9] If we adopt such a broad understanding, there is then an open question as to which of the agent's conative attitudes should be the basis of evaluation of the agent's actions. Traditionally, the answer is that it should be the agent's desires. However, decision theorists typically assume that this basis of evaluation should be the agent's preferences over the objects of choice, which, in this case, are outcomes.[10]

Like desire, binary preference, too, can be understood as a kind of conative attitude, albeit a comparative one. Accordingly, the common move made in order to understand standard decision theory as a theory of instrumental rationality is to let the preferences over the objects of choice play the role of desires, and to interpret instrumental rationality as being about acting well in the light of those preferences. In this case, the relevant preferences over the objects of choice are preferences over outcomes. According to most decision theorists, preferences form the standard of instrumental rationality.

Let me call this notion of instrumental rationality *preference-based instrumental rationality.* It requires agents to act so as to do well by their preferences over outcomes. This constitutes a departure from the traditional notion of instrumental rationality in at least one sense: While we speak of simply desiring some end, preference is a binary relation between two outcomes. We of course also sometimes speak of desiring one thing over another. Preference could thus be understood as capturing only this comparative notion of desire. Note also that, since preferences range over outcomes, outcomes play the role of ends in *preference-based instrumental rationality.* I will argue in the next chapter that this is in fact the more significant departure from the traditional conception of instrumental rationality, and the reason why this notion of instrumental rationality ultimately fails to justify transitivity.

[9] Williams (1979) arguably articulates such a broad understanding of instrumental rationality when he argues that an agent only has a reason to do x if doing x somehow advances an element in her "subjective motivational set" S. This subjective motivational set, according to Williams, could contain various different pro-attitudes, plans or commitments.

[10] As we will see in Chapter 5, in the context of uncertainty, it is also often assumed that the agent's preferences over uncertain prospects or lotteries form the standard of instrumental rationality.

The move from the traditional notion of instrumental rationality to *preference-based instrumental rationality* is very common, but often implicit, and seldom argued for. Many use desire and preference interchangeably (see, for instance, Elster (1983), Chapter 1). Others equate ends with outcomes. In this passage, for instance, Morris and Ripstein (2001) claim that decision theory requires agents to have rankings of ends:

> The traditional theory of rational choice begins with a series of simple and compelling ideas. One acts rationally insofar as one acts effectively to achieve one's ends given one's beliefs. In order to do so, those ends and beliefs must satisfy certain simple and plausible conditions: For instance, the rational agent's ends must be ordered in a ranking that is both complete and transitive. (p.1)

Yet others claim that ends and desires are different from preferences over outcomes, but still abide by *preference-based instrumental rationality.* Gauthier (1987) claims that ends may be inferred from preferences, but that preferences are basic, and that rationality is about maximizing a measure of preference:

> The theory of rational choice takes as primary [...] the relation of individual preference. [...] The theory does not analyze particular relations of preference, which are treated as ultimate data, but sets of these relations, each set representing the preferences of one individual over the pairs of realizable outcomes in a choice situation. (p.22)
>
> [I]n identifying rationality with the maximization of a measure of preference, the theory of rational choice disclaims all concern with the ends of action. Ends may be inferred from individual preferences; if the relationships among these preferences, and the manner in which they are held, satisfy the conditions of rational choice, then the theory accepts whatever ends they imply. (p.26)

Nozick (1993), too, claims that preferences are basic, and that ends and desires can be derived from them through some process of filtering or processing (p.144). Hampton (1994) provides a critique of standard decision theory that relies on interpreting decision theory in terms of *preference-based instrumental rationality.*

Crucially for us, *preference-based instrumental rationality* appears to make it easy for us to justify *maximization* as a requirement of instrumental rationality. If instrumental rationality requires us to act well in the light of our preferences over outcomes, then, provided there is a most highly ranked outcome, instrumental rationality seems to require us to take the action that leads to it. If I choose in this way, I will not frustrate any of my binary preferences.

Still, temptation problems, where an agent's preferences change temporarily, have been said to put into question whether *maximization* really is a requirement of instrumental rationality. In the following, I will present these arguments and reject them.

2.4 A Temptation Problem

Suppose that I like to stream an episode of a TV show when I take my afternoon coffee break. As my coffee break starts, at t_1, I prefer to watch only one episode, and then get back to work. But once I have watched that first episode, at time t_2, I prefer to watch another one over stopping. Once I have watched the second episode, I return to my earlier preferences and would prefer just having watched the one episode.

Let O_0 be the outcome of watching no TV: I will get all my work done, but my coffee break will be boring. Let O_1 be the outcome of watching one episode, namely that I have an interesting coffee break, and also get all my work done afterwards. O_2 is the outcome of watching two episodes: While I get to watch two episodes of an interesting show, I will not get my work done. Let $\succ$ represent strict preferences between outcomes. My preferences at the different points in time are the following:

$$t_1 : O_1 \succ O_0 \succ O_2$$

$$t_2 : O_2 \succ O_1 \succ O_0$$

$$t_3 : O_1 \succ O_0 \succ O_2$$

The dynamic decision problem I face can now be illustrated by the decision tree in Figure 2.1. The square nodes here represent choices I need to make. In each case, I can decide whether to go 'up' or 'down'.

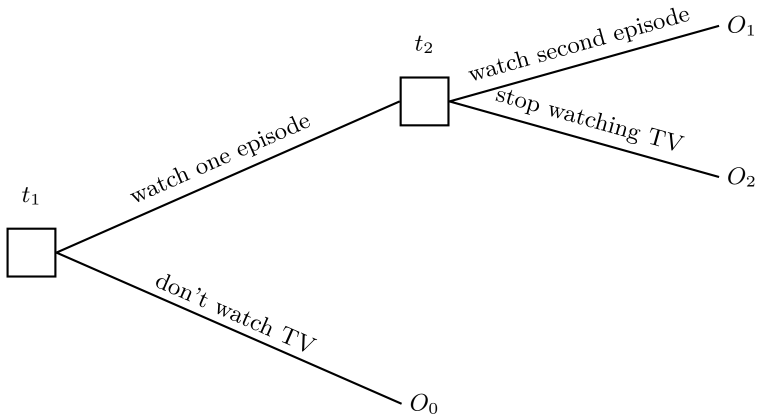


Figure 2.1: Temptation Problem

Now suppose that if I get to make a decision at time t_2, I simply choose according to my preferences between the outcomes O_1 and O_2, the two outcomes available to me then. That is, I choose to watch a second episode. In that case, I would be following *maximization.* Further suppose that I predict that I will do so at t_1, and treat this as certain. I then take myself to effectively face the choice between O_0 and O_2 at t_1. If again, I simply go with my preference over these outcomes at t_1, I will choose to not watch any TV. I will do so even though, at every point in time, I prefer watching one episode over watching no episodes.

As we said, this kind of case is often referred to as a *temptation problem.*[11] Note that unlike in traditional cases of weakness of will, this is not a case where an agent is tempted to act against her best judgement. Instead, here the agent is faced with a temporary shift in her preferences. In traditional discussions of weakness of will, debate has focused on whether and how it is even possible to act against one's own best judgement.[12] Here, the question is whether it could be rational to act against one's

[11] See, for instance, Bratman (1998). Holton (2009) argues that agents who face a tempting situation will in fact often adjust their preferences to favour giving in. Thus weakness of will turns into temptation as we understand it here.

[12] Socrates famously denied the possibility of weakness of will thus understood in Plato's *Protagoras.* See Stroud (2014)

preferences in cases where preferences have shifted temporarily.

The choice behaviour I have just described is referred to as 'sophisticated' in the literature on dynamic choice.[13] At each point in time, the agent predicts what she will rationally do in the future. And, under conditions of certainty, she chooses in accordance with what act brings about the outcome she most prefers at that point in time, out of the ones that are still available to her then. If the agent treats her prediction of future rational behaviour as certain, sophisticated choice simply follows from continued application of *maximization.*

If I follow this choice strategy in the temptation case, then I will end up not watching TV. It is rationally impossible for me to watch one episode and then not give into the temptation to watch another. However, many philosophers have thought that it must be rationally possible for agents to resist temptations of this sort. And then there must be something wrong with *maximization* as a requirement of instrumental rationality, or else resisting temptation is in conflict with instrumental rationality.

Indeed, one main source of the intuition that resisting temptation can be rational is itself instrumental in nature. And that is that the agent in these examples seems to end up worse off by her own lights. It seems like her life would go better if she had the capacity, at t_2, to not act in accordance with her temporary preferences. The best way to see that is to interpret the choice to not even watch the first episode as a kind of costly pre-commitment mechanism. I forego the first episode to bind myself not to watch another. The cost of pre-commitment, however, only seems to buy me something I could have had for free, had I only been able to resist the temptation. We may consequently want to provide an argument that claims that it is rational to act against one's preferences at the time of temptation because doing so leaves the agent ultimately better off.

Two types of such arguments can be distinguished. On the first, it is sometimes rational to resist temptation because doing so is part of the best course of action, or called for by the best deliberative strategy, by the agent's own lights. According to these arguments, the rationality of the individual action should be assessed by whether it is endorsed by the best deliberative strategy, or best course of action. Let us call these "instrumentalist two-tier arguments". On the second type of argument, resisting temptation is rational because it is the product of mutually beneficial cooperation between the agent's time slices. Let us call these "time-slice cooperation arguments".

Both of these types of argument try to show that resisting temptation is rational with appeal to instrumental rationality alone. And both types of arguments standardly presuppose *preference-based instrumental rationality.* Here I want to argue that both arguments fail under the assumption of *preference-based instrumental rationality*, but are largely redundant once the assumption is abandoned. If we are committed to *preference-based instrumental rationality*, this is good news for *maximization* as a requirement for instrumental rationality, but bad news for those who think that resisting temptation need not be irrational.

2.5 Instrumentalist Two-Tier Arguments

Instrumentalist two-tier arguments are based on the claim that agents sometimes serve their desires best if they do not, at every point in time, take their reasons directly from their desires. This is the basic insight David Gauthier (1994) provides in his "Assure and Threaten". Under *preference-based*

for an overview of the contemporary debate.

[13] For a formal discussion of this an other dynamic choice rules, see McClennen (1990). Sophistication will also be discussed again in Sections 4.2, 5.5 – 5.8, and 6.4.

instrumental rationality, the agent's preferences over outcomes play the role of desires in instrumental rationality. And then the claim is that agents sometimes do better according to their own preferences if they are not sophisticated, and do not act in accordance with *maximization* at each point in time.

Given this basic insight, Gauthier argues that instrumental rationality in fact demands that we assess not individual choices, but entire deliberative procedures by how well they serve our preferences. We then regard actions as rational if and only if they are in accordance with the best deliberative procedure — even if that procedure calls for a choice that serves the agent's preferences at the time of action less well than another. This is the sense in which he defends a 'two-tier account': More general deliberative strategies are assessed instrumentally, but individual acts are assessed according to whether they are endorsed by the deliberative strategy.

There are various worries about the two-tier nature of this account. For instance, we do seem to have a strong intuition that whether an action is instrumentally rational depends on how well it serves the agent's ends at the time of action, insofar as they are still attainable then. In fact, Bratman (1998) calls this the *standard view*. Denying it would imply that we can be moved by the 'dead hand of the past'. What options were once attainable to me but no longer are does not seem relevant anymore at the time of action. What past opportunities I had is simply another fact about the past that I cannot change anymore. Similar claims apply to the preferences I once held but no longer hold.

In our terms, the worry is that *maximization*, under the assumption of *preference-based instrumental rationality*, has a lot of intuitive appeal. For this reason, we may be tempted to say that even if a deliberative strategy that violates *maximization* better served the agent's preferences, instrumental rationality nevertheless requires *maximization*.

However, here I want to raise another, more fundamental problem for instrumentalist two-tier arguments for the rationality of resisting temptations. And that is that in cases of temptation, we cannot in fact establish that a deliberative strategy that endorses being resolute in the face of temptation really serves the agent's preferences best. And so, under the assumption of *preference-based instrumental rationality*, the argument does not get off the ground. The basic problem is that it is in the nature of temptation cases that the agent's preferences change. Under *preference-based instrumental rationality*, that means that the standard against which we evaluate deliberative procedures instrumentally changes. Later, I will argue that abandoning *preference-based instrumental rationality* does not help the argument. Save in very special circumstances, either the same problems arise, or there is no need for a two-tier account.

In the debate on temptation, it is sometimes simply taken for granted that it would be advantageous, in the agent's own lights, to be able to be resolute in the face of temptation. Most of the debate instead focuses on the question of how we can draw from this claim the conclusion that it is rational to resist temptations at the time of temptation. Take for instance the rational non-reconsideration approach to temptation introduced by Holton (2009). Holton appeals to an earlier two-tier account of the reconsideration of plans and intentions introduced by Bratman (1987). According to this account, we should assess habits of reconsideration of plans and intentions by how they affect "the agent's long-term prospects of getting what she wants." (p. 65) Then we consider an agent rational just in case her behaviour manifests 'reasonable' habits of reconsideration of her prior plans and intentions.

Bratman himself does not take this account to help us in temptation cases. However, Holton argues for such an extension. Most of Holton's argument focuses on how non-reconsideration of a resolution not to give into temptation is possible in the face of temptation and how it can keep us from in fact giving into the temptation. But he mostly seems to take it for granted that a habit of non-reconsideration would be advantageous to the agent in the face of temptation. That is, in our example, it would be advantageous if the agent had a habit of not reconsidering a previous plan to only watch one episode. This is precisely where I want to challenge the two-tier account, however.

In order to do so, I will consider a two-tier account that is more explicit about its instrumentalist basis, namely Gauthier's own. Gauthier (1994) appeals to the counterfactual consideration that the agent at each point in time thinks that she is better off going through with a resolution than she would have been had she made no resolution at all. Let us first look at the example that motivated Gauthier's argument, namely an inter-temporal Prisoner's Dilemma between two agents first described by Hume (2007/1739), III.2.5 520-521.

In this example, two farmers A and B would benefit from helping each other harvest their crop rather than doing it each on their own. However, for each, it would be even better if the other helped him harvest his field, and he would not have to reciprocate. Now we imagine that the dynamic structure of the case is such that farmer A's field is ready to harvest earlier. The farmers now face a dynamic version of a Prisoner's Dilemma, illustrated in Figure 2.2.

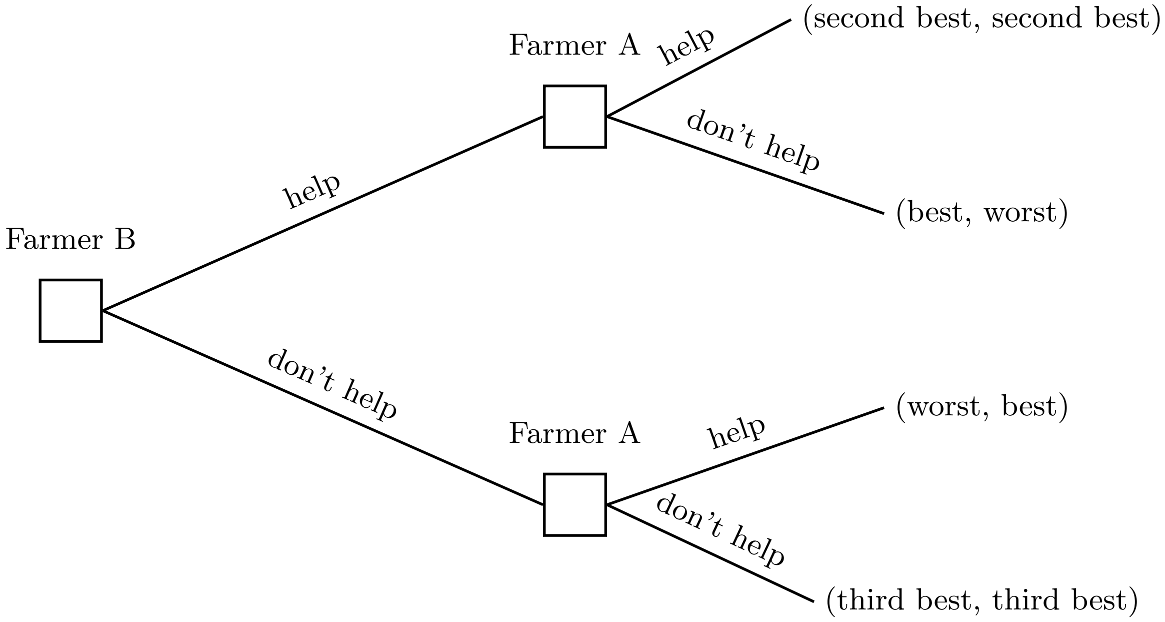


Figure 2.2: Inter-Temporal Prisoner's Dilemma

If farmer A just goes with his preferences over the outcomes still available at each point in time, then whether he was himself helped or not, he will decide not to help farmer B. In either case, he will be better off not helping. But knowing that, farmer B will not help farmer A in the first place. He would know that, whether he helps or not, the favour will not be returned. And so he himself is better off not helping. But then the farmers each end up with a worse outcome than they could have had: They each end up harvesting their fields alone, when they could have helped each other and both been better off.

The farmers could achieve the better outcome, Gauthier argues, if farmer A could make a sincere assurance in the beginning that he would help farmer B when it comes to harvesting his field, provided farmer B also helped him. If farmer B believes this, he will in fact help, in order to secure farmer A's help in return. But farmer A can only make a sincere assurance that farmer B will believe if he will take himself to have reason to follow through on the assurance when it comes to doing so. The problem here is that he will not take himself to have such reason if he takes his reasons directly from his preferences over outcomes, that is, if he adheres by *maximization.*

In this kind of case, farmer A does better with regard to his preferences if he uses a deliberative procedure that requires of him to go through with his assurance, even if it means at times not choosing the act that he prefers. If instrumental rationality is ultimately about doing well by our preferences, then, Gauthier argues, it endorses such an alternative deliberative strategy, and assessing the rationality of individual actions by how well they conform with the strategy.

In a nutshell, the specific deliberative strategy that Gauthier defends in this kind of case is the following. When it comes to following through with an assurance, or any kind of commitment or resolution, the agent should ask herself two questions. First, 'how well would I have done if I had never made any assurance or resolution'? And second, 'how well will I do if I follow through with the assurance or resolution'? If the agent judges she would do worse following through with the assurance or resolution than she would have done having never given an assurance or made a resolution at all, then she is free to just maximize. But otherwise, she should follow through with the assurance or resolution. The assurance or resolution is thus appraised against a counterfactual, the counterfactual of what would have happened had no assurance or resolution been made.

Let us assume that in the absence of an assurance, both agents are sophisticated and simply maximize at each point in time, and know this about each other. They thus know that in the absence of an assurance, they will each have to harvest their fields alone. Then the deliberative strategy Gauthier proposes provides us with the desired result in the inter-temporal Prisoner's Dilemma, provided farmer A has made the assurance. When it comes to helping farmer B, farmer A judges that he will do better following through with his assurance than he would have done never having made an assurance at all.

Interestingly, this deliberative strategy also applies to the temptation case presented in Section 2.4 above, and can be used to justify going through with resolutions to resist temptations. Suppose the agent in the beginning of the decision problem made a resolution to watch only one episode and then stop. At t_2, when it comes to following through, she considers what would have happened had she not made this resolution. Again, we assume that in the absence of a resolution, the agent is sophisticated. She would then not even have watched the first episode. That means, according to the agent's preferences at t_2, she would have done worse not having made a resolution than she would do following through with the resolution. Even then, she prefers only watching one episode to watching none. And then according to Gauthier's proposed deliberative strategy, she should follow through with the resolution.[14]

[14] While Gauthier (1996) argues in favour of extending the two-tier account to justify resolution in temptation cases, Gauthier (1997) in fact expresses some scepticism that it can be so extended. But seeing that the deliberative strategy he proposes in "Assure and Threaten" applies to temptation cases, he will need to modify the account he offers there to be able to justify such a restriction. Moreover, that modification would need to be well motivated. The reason Gauthier (1997) states for not extending the account to temptation is that there is a crucial difference between the intra-personal temptation cases and the inter-personal Prisoner's Dilemma. That difference is that the agent does not relate to herself over time as she does to other people. In the inter-personal Prisoner's Dilemma, Gauthier claims, "an agent will normally have two reasons to be resolute – one based on the role of mutually beneficial interactions in realizing whatever concerns she has, the other based on the character of her relations with at least some other persons." (p.18) Part of the reason in favour of cooperation, in the inter-temporal Prisoner's Dilemma, is that the agent views other people as 'ends in themselves'. But Gauthier claims that she does not view her previous selves in that way. However, note that Gauthier here seems to abandon our presupposition, and the presupposition he makes in 'Assure and Threaten', that resolution is to be justified in terms of instrumental rationality alone. Gauthier seems to claim that resolution can be justified only by appeal to the substantive requirement that agents treat other people as ends in themselves. And so while I agree in the following that temptation cases are crucially different from the inter-temporal Prisoner's Dilemma, the difference Gauthier himself points out is one that he cannot appeal to if he wants to offer a true instrumentalist two-tier account.

2.6 Problems with Instrumentalist Two-Tier Arguments

There is, however, a crucial difference between the temptation cases and the inter-temporal Prisoner's Dilemma, and I want to argue that this shows that Gauthier's argument is unsuccessful in the case of temptation, even if it were successful in the case of the inter-temporal Prisoner's Dilemma.[15] In the inter-temporal Prisoner's Dilemma, each farmer's preferences over the possible outcomes of the game remain constant. These constant preferences can be used as instrumental standards by which to evaluate the deliberative strategy Gauthier proposes. It is according to his stable preferences over outcomes that farmer A does better adopting a strategy whereby he can make sincere assurances and carry them out, rather than acting in accordance with his preferences at each point in time. Farmer A prefers the outcome of both farmers helping each other to the outcome of them each harvesting alone. This is the basis of the instrumentalist argument for a deliberative strategy that makes it possible to make good on an assurance.

The temptation cases are different in this respect. Here, the agent does not have constant preferences over outcomes. Gauthier claims that agents should use the deliberative procedures that best serve their goal of their lives going as well as possible. For a two-tier account to apply, we need to identify a deliberative strategy that is best by the agent's own lights. But if the agent's conception of what that goal consists in changes, we have no constant standard by which to evaluate deliberative procedures.

The deliberative procedure Gauthier proposes results in the best outcome according to the agent's preferences at t_1. But it does not lead to the best outcome according to the agent's preferences at t_2. At t_1, the agent thinks that the best course of action is one where she watches only one episode and then stops, since this leads to her most preferred outcome. But at t_2, according to her preferences, the best course of action for the whole choice problem is the one where she watches the first episode and then goes on to watch another. This would lead to her most preferred outcome. According to the agent's preferences at t_2, a deliberative procedure that endorses this course of action would be best.

Gauthier's proposed deliberative strategy can endorse making a resolution to not watch a second episode and going through with it, as we have seen. But it would not equally endorse making a resolution to watch two episodes and going through with that resolution — which is the best course of action according to the agent at t_2. At t_2, the agent would have no problem going through with such a resolution, of course. But at t_1, the agent takes it to be better to have made no resolution at all than to act in accordance with it and watch the first episode. This is because, at t_1, she prefers watching no TV over watching two episodes.

We can, however, imagine possible alternative deliberative strategies that would allow the agent to make a resolution to watch both episodes and go through with that resolution. The agent at t_2 would prefer such a deliberative strategy. Gauthier's proposed deliberative strategy is thus not the best deliberative strategy according to the agent's preferences at each point in time. It is the best deliberative strategy according to the agent's preferences at t_1. But it is not the best deliberative strategy according to the agent's preferences at t_2.

Therefore, an argument that requires an agent's deliberative strategy to be best by her own lights

[15] Interestingly, the rationality of resisting temptation is much less controversial than the rationality of going through with assurances in inter-personal cases. In fact, Hedden (2015b) argues against requirements to avoid intra-personal diachronic tragedy, e.g. by resisting temptation, by likening intra-personal diachronic tragedy to the kind of inter-personal tragedy we find in the Prisoner's Dilemma, which he thinks rationality obviously does not require us to avoid. What we find here, however, is that in one respect the rationality of resisting temptation is in fact harder to establish — at least as long as we are committed to *preference-based instrumental rationality.*

in order for it to be rational to follow it does not go through. At the time when the agent is tempted, she in fact does not think that a deliberative strategy that requires her to resist the temptation is best. And so according to such an argument, she would not be rationally required to follow it.

Gauthier (1997) proposes a different deliberative strategy specifically for the context of temptation. There he notes that often, in cases of temptation, while the agent's proximate preferences for, e.g. watching a second episode, change, the agent retains 'vanishing point' preferences that still favour watching only one episode. These vanishing point preferences are preferences about how to choose in similar situations in the future. So even while the agent is tempted, she may prefer not to give into a temptation at future points in time.[16] Let us grant that this is so in our TV consumption case. Even as I am tempted to watch another episode, I prefer that I only watch one episode at my coffee break the next day. Gauthier (1997) thinks that this makes it the case that the best deliberative strategy is one where the agent ignores his proximate preferences, but acts in accordance with the vanishing point preferences he holds at other times:

> He is able to understand that if, given proximate preferences, he chooses the action that best realizes his immediate concerns, he is deliberating in a way that may not lead him to the best realization of his overall concerns, as viewed at that or at any other time. (p.20)

It is clear that, given her vanishing point preferences, the tempted agent judges that she will do much better by adopting a deliberative strategy that will make her resist temptation at all points in time than she would do if she adopted a deliberative strategy whereby she always gives into temptation. However, that does not make it the case that the agent takes the deliberative strategy of always going with her vanishing point preferences to be *best*. In particular, a deliberative strategy whereby she can make just this one exception would be preferred by the tempted agent.

Gauthier's argument only goes through on the assumption that the agent is committed to adopting deliberative strategies that commit her to treating similar decision problems alike. However, such a requirement does not seem to follow from instrumental rationality alone. Without any desire for such consistency,[17] the agent could always formulate deliberative procedures that allow for exceptions that are indexed to a specific time or place.

At this point, we might want to make a two-tier argument at a higher level, to the effect that agents who don't allow themselves to make exceptions generally do better in life. But again, as long as the agent's shifted preferences are the standard of instrumental rationality, the best deliberative strategy at this higher level will be one that allows just this one exception to not making exceptions.

The underlying problem for both of Gauthier's accounts seems to be that we assumed that the agent's preferences provide the standard against which to evaluate a deliberative strategy. As these preferences change, the standard by which to evaluate the deliberative strategy changes.[18] This stands in the way

[16] We might even think that such vanishing point preferences are necessary for a case to even count as temptation proper. Tenenbaum (2016) treats this kind of case as relevantly different from cases where the agent has no such vanishing point preferences. However, for our purposes, the distinction turns out to make no difference. The two-tier argument fails either way.

[17] Interestingly, the main conclusions of Chapters 4 and 5 will be that exactly such a desire for consistency is needed in order to justify the core principles of standard decision theory instrumentally. Agents to whom standard decision theory can be justified will thus be the same kinds of agents for whom Gauthier's argument could be successful.

[18] I am assuming here that *preference-based instrumental rationality* is about doing well by the preferences the agent actually holds at the time of action. It is not, e.g., about doing well by all the preferences the agent has ever held, or will ever hold, or about doing well by the preferences the agent has held and will hold within some smaller window of time. By doing so, I am rejecting a temporally extended view of the agent's interests. I do so for two main reasons. The first is that such a view is ultimately implausible as a basis for instrumental rationality, for the same reasons as I reject interpreting

of the kind of two-tier account Gauthier wants to give, whereby an action is rational if and only if it is endorsed by the best deliberative procedure. At the time of temptation, the agent is not only tempted, but would endorse a deliberative procedure whereby she would give into temptation.

Two-tier accounts shield the tempted agent's actions from being evaluated in terms of her shifted preferences directly. It may seem like this can protect the agent from giving into temporary shifts in preference. However, given *preference-based instrumental rationality*, the shifted preferences reappear at the higher level of deliberative strategies. Even if we grant that they need not dictate individual action, they are still the standard against which deliberative strategies themselves are evaluated. For this reason, as long as preferences are the standard of instrumental rationality, it seems like there is no hope for any instrumentalist two-tier account to provide an argument in favour of resisting temptation. And this is also why going to ever higher levels of assessment will not help either.

Before considering what may happen if we give up the assumption of *preference-based instrumental rationality*, let me consider the other prominent argument for why it may be instrumentally rational to violate *maximization* in temptation problems.

2.7 Time-Slice Cooperation Arguments

Edward McClennen (1998) offers a treatment of temptation cases that is more explicit about the changing nature of the agent's preferences, which makes it impossible for us to judge the benefits of a deliberative procedure against a single set of preferences. He still thinks an appropriate, unchanging instrumental standard for this context can be formulated, however. His instrumentalist argument is based on intra-personal optimality instead, a standard he had already advocated in McClennen (1990) for the kind of dynamic choice problem discussed in Section 5.4.[19]

At first sight, his account may look like another two-tier account. The deliberative strategy McClennen defends as rationally called for under many circumstances is resolution.[20] Let a plan be a set of choices, one for each decision node the agent could find herself at in a given decision tree. Under certainty, each plan has one outcome associated with it. A resolute agent considers which plan or plans she prefers most at the outset, and then simply carries out that plan, or one of those plans. McClennen thinks that there are instrumental advantages to resolution whenever it makes possible a series of choices that is judged at least as good or better by the agent at each point in time in the decision problem, than the alternative where she is sophisticated. That is, resolution can be justified by appealing to what we may think of as Pareto improvements between an agent's 'time slices': Resolution leaves some time slices better off and no time slice worse off.

McClennen's appeal to Pareto optimality between time slices as a two-tier account in Section 2.7. The problem with such an account is that it would make it rationally non-optional to give more or less weight to one's past and future preferences. But it seems like caring about satisfying past and future preferences is just another desire an agent may or may not have. For an agent who truly does not care about her past or future preferences, it seems implausible to say that instrumental rationality requires her to cater to them. Note that this leaves open the possibility that there are non-instrumental reasons to care about one's past and future preferences, as well as the possibility of thinking of the agent's *well-being* as temporally extended. The second reason for not taking into account past and future preferences when thinking about the standard of instrumental rationality is that my argument from Section 2.9 would apply to such an account just as well. If the agent's preferences at the different points in time during the temptation problem all form part of the standard of instrumental rationality, then either the agent actually has constant interests over time, or he doesn't. If he does, then we are not in need of a two-tier argument. And if he doesn't, the two tier argument doesn't work, for the reasons just given.

[19] Intra-personal optimality had already been discussed in the economic literature as a standard in the context of changing preferences. See Peleg and Yaari (1973).

[20] See also McClennen (1990), p. 156 ff., and Sections 4.2, 5.8, and 6.4.

Resolution in the temptation cases above indeed yields such an intra-personal Pareto improvement. If the agent makes a resolution to only watch one episode and does not give into temptation, she ends up with O_1. If, instead, she is sophisticated and acts according to her preferences at each point in time, she ends up with O_0, as we have seen above. But at each point in time in the dynamic choice problem, she prefers O_1 to O_0. And so the resolute strategy is superior according to McClennen's criterion. And in fact, no further Pareto improvements are possible here, since there is no other outcome that is judged better by the agent at each point in time.[21]

The problem with interpreting McClennen's argument as a two-tier argument is that inter-temporal optimality is implausible as a standard of instrumental rationality. This is because an agent need not care about her preferences at different points in time. But treating inter-temporal optimality as a standard of instrumental rationality would make it non-optional for her to cater to her past and future preferences. A requirement to cater to one's past or future preferences even if one does not care about them does not sound like a requirement of *instrumental* rationality (even if it may be a non-instrumental requirement of rationality).[22] And McClennen himself claims to be in the business of establishing requirements of instrumental rationality. Under *preference-based instrumental rationality* as we understand it, if the agent did care about achieving inter-temporal optimality in a way that is relevant for instrumental rationality, it seems like she would have ranked the Pareto optimal outcome most highly in her preferences. Given that the agent does not rank resisting temptation most highly at the time of temptation, she thus does not sufficiently care about achieving intra-personal, inter-temporal Pareto optimality.

McClennen's appeal to optimality is thus not best understood as part of a two-tier argument. Instead, the best way to interpret his appeal to optimality is in analogy with the role of Pareto optimality in inter-personal choice problems like the Prisoner's Dilemma. In those games, there is no agent whose end it is to achieve Pareto improvements. It is simply the case that achieving a Pareto improvement serves both agent's ends. This provides the basis for authors like Gauthier to argue for the rationality of decision rules that make cooperation possible. Each agent has a reason to do her part in making cooperation possible, because each agent stands to gain from it.

McClennen suggests that analogously, in the temptation cases, the agent's 'time slices' can engage in mutually beneficial cooperation. In fact, Ainslie (1992) similarly suggests that willpower in the face of the preference reversals caused by hyperbolic discounting is the result of a kind of intra-personal cooperation. Adopting a choice rule that makes such cooperation possible is advantageous for each time slice. McClennen even goes on to suggest that for a full account of the rationality of resolute choice, we need to specify what would be a fair division of the cooperative surplus between time slices — just as

[21] Note, however, that there may be other temptation cases where no such intra-personal Pareto improvements are possible. Perhaps, for instance, I know that at 4pm this afternoon, no matter what I do, I will be tempted to have a Mohnstriezel that I would now prefer to forego. In this case, there is a clear conflict of interest between my time slices. Since no costly commitment devices can keep me from giving into this temptation, there is also no chance for the inter-temporal Pareto improvement of foregoing the costs of pre-commitment. We might think that it is already implausible that McClennen's account makes the rationality of resisting temptation dependent on whether there are possibilities for intra-personal Pareto improvement, e.g., possibilities for costly pre-commitment that the agent could forego.

[22] It is sometimes assumed or argued in decision theoretic literature that choosing rationally consists in choosing well for your future self. Jeffrey (1965/1983) appeals to this idea when arguing for his version of evidential decision theory. Briggs (2010) uses it to analyze various decision theoretic paradoxes. And LA Paul (2015) presupposes this when she argues that rational choice is impossible when we can't know what our future attitudes will be. Of course most of us care to some extent how we will view our decisions in the future. But first, this is rarely all that matters for us. Suppose I predict now that I will spend less time with my friends if I have a child, but I also predict that I will no longer care as much about spending time with my friends. If I care a lot about my friends now, then the latter prediction may give me *more*, not less reason not to have a child. And second, it cannot be a requirement of instrumental rationality that we care about our future attitudes. Instrumental rationality, I take it, is about doing well by the ends we actually hold at the time of decision.

we need an account of fair division in the inter-personal case.

2.8 Problems with Time-Slice Cooperation Arguments

Regardless of the merits of the argument in the inter-personal case, I think that this analogy fails. McClennen leaves it somewhat vague what time slices are and how they relate to the agent. But however we think of them, the analogy to inter-personal cooperation is suspect.[23]

On one end of the spectrum, we could think of the time slices as separate agents that exist in succession (but presumably retaining memory of resolutions made by earlier time slices). Carrying out a resolute choice strategy now requires different agents to do their part: One needs to form a resolution, and the other ones need to carry it out. The problem on this interpretation, apart from the implausible picture of agency it paints,[24] is that the time slice at t_2 whose turn it is to resist the temptation is asked to act on a resolution that she did not make herself. She never made any assurance to the time slice at t_1 that she would resist the temptation, and had no say in the formation of the resolution. She could not have done so, since she was not around at the earlier points in time.[25]

And so if this case resembles a case of inter-personal cooperation, it resembles one where a cooperative scheme is forced on an agent. In the farmer case, suppose that farmer A and B have not communicated at all. Farmer A harvests half his field alone and then takes a break. When he comes back, farmer B has harvested the rest of the field for him, and farmer A benefits from the field having been harvested faster. Even if farmer A knows that farmer B would only have done this had he expected farmer A to return the favour, it does not seem instrumentally irrational of farmer A not to return the favour. It might be nice to do so, or even called for by some social norms. But unless farmer A cares about these social norms or about being nice, instrumental rationality seems to in fact require farmer A to not help his neighbour in return.

The underlying problem here seems to be that if one thinks of time slices as separate agents that never exist simultaneously, there is no way in which the time slices can decide on a resolution together. The time slice that has to carry out the resolution does not exist yet when the resolution is made. But unless that time slice sees the resolution as a commitment she herself made, it seems implausible that instrumental rationality could require her to go through with it.[26]

On the other end of the spectrum, we could think of time slices as different stages of the same agent. In the temptation cases, this same agent merely changes her preferences over time. But in this kind of

[23] In the literature on personal identity, the question of what it takes for a person to persist over time is often put in terms of what relations must hold between a person's time slices for her to persist as the same person over time. Parfit (1984) famously held that no facts in the world could guarantee that two time slices are the same person, and that the most that can be said is that there are varying degrees of psychological continuity and connectedness between them. The question of what it takes to persist as an agent is arguably different from the question of what it takes to persist as a person over time. In any case, I here don't want to take a stance on either question, but argue that McClennen's analogy does not work under any way of thinking about time-slice agency.

[24] In fact, McClennen (1990) himself is sceptical of thinking of time slices as ontologically separate in the same way different agents are (p.15).

[25] Bratman (1995) similarly objects to appeals to intra-personal optimality on the basis that the earlier time slice is not around anymore once the later time slice gets to make a choice. The concept of cooperating with the dead, as it were, seems odd. I take that objection not to be entirely decisive. if we hold a preference satisfaction view of benefit, then we may think that agents can be benefitted even if they don't know they are, or even after they are dead. In that case, we might think that it's not entirely implausible that our best account of inter-personal cooperation requires agents to do their part even once their cooperative partner is 'not around anymore'.

[26] Cases where a very large benefit is possible, that requires the cooperation of many people, may put pressure on the intuition that explicit agreement is necessary for instrumental rationality to demand cooperation. I want to take no stance on this here, but only note that this is no such case.

case, we usually simply assume that the new preferences override the old preferences, and there is no reason for the agent to still act on preferences she does not hold anymore. In cooperation, too, there seems to be no reason to make good on an assurance if doing so would not benefit the agent you are cooperating with anymore, due to a shift in her preferences.

For instance, in the farmer case, suppose farmer A secured farmer B's help with an assurance. But just before it comes to reciprocating, farmer B changes his preferences such that he now prefers harvesting alone after all. Perhaps he took a sudden dislike to farmer A. It seems implausible that in this kind of case, there is anything to be said for farmer A helping farmer B. In fact it would be bizarre for farmer A to still impose his help.[27] Likewise, it seems, in the case where the tempted agent is cooperating with herself, there is nothing to be said for catering to the agent's earlier preferences once they are overridden.

The best way to think about time slices may lie somewhere in the middle. But two requirements would need to be met in order for the argument to resemble inter-personal cooperation like in the original inter-temporal Prisoner's Dilemma. First, it would need to be the case that time slices are separate agents in the sense that the preferences of later time slices do not override the preferences of earlier time slices. And second, the time slices need to belong to the same agent in the sense that a later time slice recognizes a resolution made by an earlier time slice as her own. I don't see how these two requirements could plausibly be met together.

2.9 Giving up Preference-Based Instrumental Rationality

The previous arguments were meant to show that the prominent instrumentalist arguments for the rationality of resisting temptations fail. Two-tier accounts fail because in temptation cases the standard by which to evaluate deliberative procedures shifts. And no plausible account of mutually beneficial cooperation between time slices of an agent can be given.

However, note that my argument relied on the assumption of *preference-based instrumental rationality.* But this assumption may well be false. In fact, the next chapter will argue that there is independent reason for thinking that it is false. As we will see, the assumption makes it impossible to provide an instrumental justification for transitivity. Still, giving up *preference-based instrumental rationality* does not help those who want to make instrumentalist two-tier arguments for resisting temptations.

Chapter 4 will consider in more detail what an alternative to *preference-based instrumental rationality* may look like. But there are a few observations we can already make. First, if there is to be any hope of a decision theory that is formulated in terms of preferences to serve as a theory of instrumental rationality, then preferences should at least normally or ideally be responsive to the true standard of instrumental rationality, or provide some kind of representation of it. Let us vaguely call this true standard the agent's desires for now. Second, giving up *preference-based instrumental rationality* opens up the possibility that preferences incorrectly represent this true standard of instrumental rationality.

We may suspect that this is what is going on in temptation cases: Under the influence of some temptation, the agent's shifted preferences diverge from her underlying, true desires. As I want to argue here, however, conceding this does not help those who want to make the instrumentalist arguments I considered here. We can distinguish three exhaustive possibilities of what may be going on in temptation

[27] An exception might be a case where farmer B told farmer A to help him in return no matter what — that is, even if he later says he wants no help. But still it seems like, in such a case, the later self's preferences override this earlier wish, unless the later self is not in full possession of his rational faculties. But we are assuming that this is not the case in temptation problems.

cases, where preferences temporarily change.

First, it could be that while the agent's preferences as we describe them in some temptation case may not correctly capture her desires, there is no stable ranking of the options in the temptation case that would correctly capture the agent's desires at every point in time. This would be evidence of the agent's underlying desires in fact shifting significantly. The temptation does not disappear once preferences are altered to correctly capture the agent's desires. If this is so, all the problems we discussed in the foregoing still arise.

Another possibility is that the only preferences that would correctly capture the agent's desires throughout a temptation problem would be stable ones. This would be evidence of the agent having stable underlying desires after all. It may well be that appeal to stable underlying desires that speak in favour of resisting temptation is what Holton has in mind as the basis for his argument that a habit of non-reconsideration would be rational. It is also suggested by Sarah Paul (2015) who claims that the stable, more long-term preferences an agent has before and after being tempted have a better claim to 'speak for the agent' (even at the time when she is tempted). Gauthier's (1997) argument that the agent should act on her 'vanishing point' preferences may also in part have been motivated by this intuition. The fact that the preference reversal in temptation cases is only temporary could be seen as evidence that tempted agents never stop having the goal of being temperate, but are only momentarily confused about what they really want.[28]

However, in this case, it seems like we don't need the instrumentalist arguments we have been considering in the foregoing anymore. What is instrumentally rational is to do well by one's desires. If the agent's desires, all the way through, uniquely support only watching one episode, then even as the agent is tempted to watch another episode, instrumental rationality requires her to not watch the second episode. This is so for straightforward reasons. Perhaps not reconsidering a resolution to only watch one episode is a good way for an agent to conform to this requirement. But it seems like as long as the agent in fact refrains from watching the second episode, however she manages to do so, she is instrumentally rational.

If these are the only two possibilities, then there is no use for the arguments we considered here. Either they fail, or they are redundant. Where would this leave *maximization*? *Maximization* remains untouched if we stick with *preference-based instrumental rationality.* If we reject it, as we have said, there is a possibility that an agent's preferences may not accurately reflect her underlying desires. So at best, *maximization* is defensible only as a conditional principle: An agent ought to maximize *if* her preferences accurately capture her underlying desires.[29]

If we stick to *preference-based instrumental rationality*, we cannot give instrumentalist arguments for resisting temptation. Only if we abandon it will we be able to give an instrumental argument for resisting temptations. We can give such an argument if we can make a convincing case that the agent's true underlying desires support resisting the temptation after all. The cost is that we give up *preference-based instrumental rationality*, which we have seen means we can at best defend a restricted version of *maximization.* As it stands, however, introducing a condition that the agent's preferences are true representations of her desires does not seem like a very high cost to pay, and I expect that many decision theorists will be happy to take it on board. The next two chapters will show, however, that the

[28] This is also suggested by Ainslie's (2001) suggestion that the preference reversals characteristic of hyperbolic discounting are caused by a kind of perceptual illusion.

[29] Section 4.6 in Chapter 4 will show that, in fact, not even that conditional version is defensible, given a plausible non-uniqueness thesis.

abandonment of *preference-based instrumental rationality* has more far-reaching consequences.

One such consequence needs to be mentioned here already. And that is that abandoning *preference-based instrumental rationality* may lead to a third possibility of what is going on in temptation cases. If we allow for several different preference relations to equally represent an agent's desires at a particular point in time, the following might happen. It could be that there is at least one stable preference order that would correctly capture the agent's desires at every point in time, but that at any point in time, several different preference relations would accurately capture the agent's underlying desires. The tempted agent has shifting preferences, but she could have stable preferences that would capture her desires correctly. This is consistent both with the agent having stable desires, and with her having desires that shift only slightly over time.

This is a possibility we will only be able to do full justice to in Chapter 4. I will argue there that in these circumstances, an instrumentalist two-tier argument may give the agent reason to stick with one of the preference orders that represent her desires correctly at every point in time throughout, or to act as if she did. In fact, we will see that the availability of this kind of instrumentalist argument is crucial for what I take to be the best attempt at an instrumentalist justification for the core principles of standard decision theory.

However, in the case of temptation, this seems to me to be a special case, and only some real life temptation cases will be accurately described by this analysis. And furthermore, as we will see, the question of how an agent should act in these circumstances is orthogonal to the question of whether *maximization* holds. And so while this may be a vindication of two-tier arguments in some circumstances, it is not a vindication of two-tier arguments in favour of resisting temptation more generally, or of two-tier arguments against *maximization*.

2.10 Conclusions

Maximization is one of the central requirements of standard decision theory in the context of certainty. It appears to be easily justified under the assumption of *preference-based instrumental rationality*. This chapter considered arguments to the effect that instrumental rationality in fact requires us to violate *maximization* in cases where our preferences shift temporarily. I argued that these arguments fail under the assumption of *preference-based instrumental rationality*, and, save for a special case, are redundant if we abandon this assumption.[30]

For temptation cases, the implication is that an important kind of argument for the rationality of resisting temptations fails. But our last considerations suggest another way of responding to the temptation cases that is still instrumentalist in character. And that is to show that it will often be the case in temptation cases that the agent's desires in fact support resisting the temptation. The agent's momentary preferences merely do not reflect that.

Two things need to be noted about such a response: The first is that if we take this route, we can no longer give an argument that instrumental rationality *requires* that agents resist temptation, unless having the desire in question is itself required for instrumental reasons rooted in the agent's other ends.[31]

[30] This is not to preclude that other types of cases may in fact give us reason to abandon *maximization*. Gauthier's inter-temporal Prisoner's Dilemma may be such a case, and so could the toxin puzzle presented by Kavka (1983). As Tenenbaum and Raffman (2012) argue, dynamic choice problems involving vague ends may also lead us to abandon *maximization*. In fact, the next chapters will take up some of their argument to show that no instrumental justification can be given for *maximization* once we abandon *preference-based instrumental rationality*.

[31] This is, in fact, the route Bratman (2017) takes. He argues that being self-governed requires an agent to desire to

After all, instrumental rationality cannot demand that the agent have any particular desires. It hence cannot require the agent to have desires that support resisting the temptation, even at the time of temptation. The best we can do is to argue that agents ordinarily have desires that support resisting temptations in a wide variety of cases.

The desires we appeal to here could be specific to each temptation case, or they could be desires that speak in favour of stable preferences more generally. The next two chapters will argue that instrumental arguments for the transitivity of preference only go through for agents who in fact have a general desire that favours stability of preference. While I will identify the best case that can be made for such a desire, I will argue that we are not rationally required to have it, at least not by instrumental rationality.

For standard decision theory, the implication of my argument in this chapter is that *maximization* remains unchallenged under the assumption of *preference-based instrumental rationality.* To make instrumental arguments in favour of resisting temptation possible, *preference-based instrumental rationality* needs to be given up. In that case, *maximization* can at best only hold conditionally: An agent ought to maximize if her preferences correctly capture her desires. The next chapters will show that abandoning *preference-based instrumental rationality* has more far-reaching consequences.

be self-governed, and this desire could support an instrumental argument in favour of resisting temptation. Moreover, there is, for what Bratman calls 'planning agents', rational pressure to be self-governed. And at least part of this rational pressure is instrumental in nature.

Chapter 3

Preference-Based Instrumental Rationality and the Money Pump Argument

3.1 Introduction

Most decision theorists want to defend decision theory as a theory of instrumental rationality. As we have argued, they typically view the agent's preferences over fully specified outcomes as the standard of instrumental rationality: Actions are rational if they serve the agent's preferences well. *Preference-based instrumental rationality* makes it apparently easy for us to justify *maximization*, one of the central requirements of standard decision theory in the context of certainty. If instrumental rationality consists in doing well by my preferences, then it seems like I shouldn't choose an action to which another one would have been preferred.

In the last chapter, I argued that one common objection to *maximization* fails under the assumption of *preference-based instrumental rationality.* We cannot argue that in the context of temptation, it is sometimes instrumentally rational to act counter-preferentially. In fact, if we want to give an instrumentalist argument for the rationality of resisting temptation, then we need to abandon *preference-based instrumental rationality.* This chapter shows that we are justified in doing so for independent reasons.

I here turn to the other core requirement of standard decision theory in the context of certainty, namely the requirement that our preferences should form a weak ordering over outcomes. One important part of this requirement is that preferences should be transitive. On the face of it, this looks like a non-instrumental requirement of rationality. It is a requirement on what kinds of preferences an agent may hold. But instrumental rationality was supposed to be silent on what ends an agent may have. In fact, Hampton (1994) rejects the Humean interpretation of standard decision theory on that basis.

There is, however, one prominent instrumentalist defence of transitivity. I will here consider whether this defence can establish standard decision theory as a theory of instrumental rationality after all. This defence comes in the form of money pump arguments. These are arguments to the effect that agents with intransitive preferences will accept series of choices that have them pay for ending up with what they started with. This is argued to be instrumentally irrational.

This chapter shows that money pump arguments do not go through on the assumption of *preference-based instrumental rationality.* I argue that for this and other reasons, there is a decisive case for abandoning this notion of instrumental rationality. Its basic flaw, I claim, is that the objects of the preferences that guide choice in standard decision theory in the context of certainty are fully specified outcomes. Yet, the objects of our basic desires are more simple states of affairs.

The next chapter will show that abandoning *preference-based instrumental rationality* has far-reaching consequences. In particular, abandoning it again puts into question *maximization.* I will argue that the best instrumentalist case that can be made for the core requirements of standard decision theory under certainty only applies to agents who already have a specific kind of desire.

3.2 Cyclical Preferences

Intransitive preferences are both wide-spread,[1] and often appear sensible. Two types of cases are commonly used in order to motivate the idea that intransitive preferences may sometimes not be irrational, or may even be called for. Both of these are cases of cyclical preferences, that is, preferences that form a loop of strict preferences over a set of outcomes. In one kind of case, the possible outcomes of the different actions available to the agent differ in various different dimensions the agent cares about. And in the other kind of case, the outcomes of the actions available to the agent are in some respects seemingly indistinguishable to her.

To start with the first kind of case, suppose I am looking for an apartment. Three apartments are available for the same rent, which I can afford. They differ only in terms of their size, their views, and the length of the commute I would have if I lived in the apartment. All three of these are factors that I care about.

Apartment A: 40 m^2 large; view onto a garden; 5 minute commute.

Apartment B: 70 m^2 large; view onto the skyline, lake and woods; 60 minute commute.

Apartment C: 100 m^2 large; view onto the brick wall of the building next door; 30 minute commute.

When it comes to choosing where to live, my pair-wise preferences over the outcomes of living in each of these apartments (denoted by 'Apartment A', 'Apartment B', 'Apartment C') may well be cyclical, in the following way:

$$\begin{gathered}\text{Apartment B} \succ \text{Apartment A}\\ \text{Apartment C} \succ \text{Apartment B}\\ \text{Apartment A} \succ \text{Apartment C}\end{gathered}$$

What may make these preferences seem defensible is that I can give the following explanation of my preferences: I prefer Apartment B over Apartment A because Apartment B is larger and has such a lovely view, and this outweighs the fact that it has a longer commute. I prefer Apartment C over Apartment B because Apartment C is even larger, and has a shorter commute, and this outweighs the fact that it does not have a good view. And I prefer Apartment A over Apartment C, because it has an even shorter commute, and a better view, and this outweighs the fact that it is smaller.[2]

[1] See, for instance, Loomes et al. (1991).

[2] Also see Rabinowicz (2000) on the claim that multi-dimensional decision contexts are perfect breeding ground for cyclical preferences.

The second kind of case is best illustrated with the Puzzle of the Self-Torturer, first introduced by Quinn (1990). Suppose somebody straps a device to your arm that causes you pain with electric shocks. The device has 1,000 different settings. At the first setting, it causes you no pain. At the highest, the pain is excruciating. However, adjacent settings differ so little in their electric current that you do not feel a difference in pain between them when experienced subsequently.

Now suppose you are offered $10,000 in exchange for each setting you are willing to go up. And so each setting of the device is associated with an amount of money. Let $S_1, S_2, S_3, ...S_{1000}$ be the outcomes of ending up with the level of pain and amount of money associated with the 1,000 different settings. Now it seems reasonable to have the following, cyclical preferences over these outcomes:

$$S_1 \prec S_2 \prec S_3 \prec ... \prec S_{1000} \prec S_1$$

Out of two adjacent settings, you always prefer the higher one. After all, you cannot detect a difference in pain between them when experienced subsequently,[3] and $10,000 is a substantial amount of money. However, when you consider the highest setting, you find the amount of pain so excruciating that you would gladly forego the fortune associated with it in order to be pain-free at the lowest setting.

According to *weak ordering*, the preferences we described in both of these cases are irrational, because they are intransitive in virtue of being cyclical. But can we give an instrumental justification for *weak ordering*? Can we show that agents who have the preferences we just described are instrumentally irrational? There is a famous argument that aims to show that agents with cyclical preferences can be pragmatically criticized, namely the ‘money pump argument’. This argument can thus be understood as trying to provide an instrumental justification for at least acyclicity, which is weaker than transitivity.[4]

3.3 The Money Pump Argument

The money pump argument was first formulated by Davidson et al. (1955), but goes back to ideas in Ramsey (1928/1950). We can apply it to our examples. To start with the first, suppose a rental agency gives me the chance to choose between Apartment A and Apartment B. Choosing according to my preference between these two apartments, I go with Apartment B. But then the agency offers me the opportunity to switch to Apartment C instead. Again choosing in accordance with my preference between Apartments B and C, I choose C. Now suppose I get offered the chance to switch to Apartment A, in exchange for a small fee, say $25.

I always prefer more money to less, at least in my current circumstances. But seeing that I have a strict preference for A over C, I probably still prefer A, even when I have to pay $25 for it. If not, there will be a small enough positive amount of money ϵ that I will be willing to pay. My preferences are thus:

$$\text{Apartment A} \prec \text{Apartment B} \prec \text{Apartment C} \prec \text{Apartment A - } \epsilon$$

If I follow *maximization* at every point in time, I will end up with Apartment A having lost $25. But I could have had Apartment A without losing that money, if I had only chosen Apartment A right away.

[3] Voorhoeve and Binmore (2006) and Arntzenius and McCarthy (1997) argue that the settings cannot literally be indistinguishable to the agent. All that matters for us, however, is that even if they are right, the self-torturer’s preferences seem intuitively reasonable. This could be because two adjacent outcomes are still subjectively indistinguishable when directly compared, or because any difference in pain will be expected to be tiny. And so, even if these authors are right, we are in need of an argument for why the self-torturer should not have cyclical preferences.

[4] I will focus on money pump arguments in favour of acyclicity here. However, money pump arguments in favour of the stronger condition of transitivity have also been offered in the literature. See, for instance, Gustafsson (2010).

Ending up with Apartment A having payed \$25 seems instrumentally criticizable. Moreover, the rental agency could potentially repeat offering me this series of swaps, effectively turning me into a 'money pump'.[5]

The same fate could meet you in the second example. Suppose you are offered the chance to go up by one setting every week, in exchange for the \$10,000. Going with your binary preference between two adjacent settings, you should always go up by one setting, all the way to the highest setting. This way, you would turn yourself into a self-torturer, giving the problem its name. Now suppose somebody offers you the chance to go back to the lowest setting, in exchange for giving up your entire fortune, plus an additional \$25 (or a small enough amount of money ϵ). You gladly accept. But again, you could have been pain-free for less money, making you apparently instrumentally criticizable. Moreover, the cycle could be repeated, turning the self-torturer into a money pump.

Neither of these arguments depend on the good that is lost being money. We could simply replace money here with small quantities of some other good that the agent desires. For instance, in the apartment hunting case, the cost of switching back to Apartment A might be the effort and time it takes to go back to the rental agency to sign a new agreement. Or it could be the embarrassment of changing my mind once again. The evil scientist who attached the pain device to your arm might ask you to run one lap in a human-sized hamster wheel in exchange for going back to the first setting. Or she might ask for the secret ingredient in your margarita popsicle recipe. The general problem the money pump arguments point to is supposed to be that an agent may end up losing some of a good she desires without getting anything else in return — be it money, her free time, her dignity, or her well-kept secrets. In the following, where we speak of money, this is is just a place-holder for any good the agent may desire.

Susceptibility to money pumping is widely held to be a good argument in favour of the acyclicity of preference. However, as I now want to argue, this argument does not work under the assumption of *preference-based instrumental rationality.*

3.4 Preference-Based Instrumental Rationality and the Money Pump Argument

Money pump arguments try to give an instrumental justification for why agents should have acyclical preferences. What kind of instrumental argument do they provide? McClennen (1990) characterizes pragmatic or instrumental arguments for principles of choice as follows:

> It can be argued that if the agent's preference and choice behaviour fails to satisfy one or the other of these principles, it will be possible to place him in a situation in which he will choose in a pragmatically indefensible manner. More specifically, the argument is that the agent will fail to achieve his intended objective or will fail to maximize with regard to his own preferences with respect to outcomes. [... A] principle of choice is valid if failure to adhere to it would result in choice of means insufficient to desired ends - in the agent pursuing his objectives less effectively than he could have under the circumstances in question. (p.4)

[5] Tenenbaum (2014a) holds that being money pumped in fact only becomes rationally problematic if it is repeated. In most of the following, I will assume that it is already problematic when it only happens once. However, if anything, the critical arguments I am going to make regarding money pump arguments go through more easily if being money pumped only once is not problematic.

McClennen understands appeal to money pumps as trying to provide this kind of argument in favour of acyclicity. However, this passage equivocates between two kinds of arguments. On the one hand, we may want to argue that agents who have cyclical preferences fail to maximize with regard to their own preferences. I will return to this argument below. On the other hand, we could argue that agents who have cyclical preferences pursue their objectives less effectively than they could have under the circumstances in question. What seems to be implied in this second kind of argument is that, if the agent did not have cyclical preferences, then the agent's objectives would be better served. The first argument does not rely on such a claim.

Neither argument is open to us under the assumption of *preference-based instrumental rationality*. Let us first look at the second type of argument. According to *preference-based instrumental rationality*, instrumental rationality consists in doing well by your preferences over outcomes. Transitivity and acyclicity are principles about what preferences an agent may hold. Essentially, what the second type of argument would now need to establish is that having different preferences would serve your preferences better. According to *preference-based instrumental rationality*, both instances of 'preference' here refer to the same preferences, namely the agent's preferences over the outcomes available in the series of choices. It is those preferences that are cyclical, but, according to *preference-based instrumental rationality*, it is also those preferences that are the standard of instrumental rationality.

When could having different preferences over outcomes serve your preferences over outcomes better? This appears to be only so if having different preferences comes with autonomous benefits, in terms of the preferences you currently hold. For instance, if an evil demon were to severely punish you for having the preferences you have, then it serves your preferences as they are now to have different preferences. Is susceptibility to being money pumped like this? To say so, we would have to show that being money pumped is bad in terms of the agent's cyclical preferences over outcomes. That is, we have to show that having cyclical preferences leads to an *outcome* that is bad in terms of those cyclical preferences, and that adopting different, acyclical preferences would lead to a better outcome according to those cyclical preferences.[6] Unfortunately, *preference-based instrumental rationality* does not allow us to say so.

First, many critics of intransitive preferences claim that cyclical preferences mean that there is no outcome that it would be rational for the agent to choose precisely because cyclical preferences do not pick out any outcome as 'best'. In fact this is exactly what the first argument, which we are going to consider below, relies on. But if we believe that, it would also seem like those preferences can also not act as a standard of what alternative preference relation may serve the agent better. And so, if the second argument has any hope of going through, the first can't.

Second, *preference-based instrumental rationality* keeps us from saying that having cyclical preferences leads to an *outcome* that is bad in terms of those cyclical preferences for independent reasons. To see that, note that money pump arguments exploit the following fact about the preferences of agents with cyclical preferences who prefer more money to less. If an agent has cyclical strict preferences over some options, then she also has cyclical preferences over a set of options that includes one of the original options with some small amount of money, or some other good, deducted. In the apartment example, if I have these cyclical preferences:

[6] One may respond here that preferences over general features of outcomes, such as preferences for having more money, could explain why it would be better to have different preferences over outcomes. But this would already go beyond *preference-based instrumental rationality* as we understand it here. *Preference-based instrumental rationality* takes preferences over full *outcomes* to be basic. In fact, appealing to conative attitudes over general features of outcomes is precisely what the desire-based alternative we introduce below does.

$$\text{Apartment A} \prec \text{Apartment B} \prec \text{Apartment C} \prec \text{Apartment A}$$

I also have these cyclical preferences:

$$\text{Apartment A} \prec \text{Apartment B} \prec \text{Apartment C} \prec \text{Apartment A - } \epsilon \prec \text{Apartment A}$$

Now intuitively, it seems like, when I am offered a choice between A, B and C, be it in a single choice or successively, I am permitted to choose any. That is, it does not matter which of my binary preferences between outcomes I frustrate. In fact, if there is any outcome that it would be rational in terms of my cyclical preferences to end up with, then each of these choices should be permitted. After all, it was arbitrary that the series of choices started with Apartment A. Given my cyclical preferences, somebody could also try to money pump me starting with Apartment C. If we want to use only facts about my preferences over outcomes to determine which outcomes it would be rational for me to end up with, as *preference-based instrumental rationality* demands, then we cannot treat outcomes A, B, and C differently.

However, to make the pragmatic argument we are considering here, it needs to be instrumentally irrational, in terms of the agent's cyclical preferences, to end up with A - ϵ. There thus needs to be a difference between the cyclical preferences between A, B, and C, and the set of cyclical preferences between A, B, C and A - ϵ. But *preference-based instrumental rationality* does not allow us to say so.

According to *preference-based instrumental rationality*, my preferences between outcomes are basic, and my actions are judged by how well they serve my preferences. A, B, C and A $-\epsilon$ in our example are all shorthand for different outcomes that involve me having some apartment and a particular amount of money. If all we can go by in judging the instrumental rationality of an action are preferences over outcomes, there is no reason to suppose that our preference of A over A $-\epsilon$ is any different from our preference of A over C. There would be no reason to suppose that frustrating the first preference would constitute an instrumental irrationality, and frustrating the second would not. I thus conclude that the second kind of pragmatic argument fails.

Still, one can try and make the first kind of argument mentioned by McClennen. Even if we cannot argue that not having cyclical preferences would serve our preferences better, it may still be the case that an agent with cyclical preferences "will fail to maximize with regard to his own preferences with respect to outcomes". Suppose that failing to maximize one's preferences with respect to outcomes is instrumentally irrational. Perhaps the argument is, then, that the agent's cyclical preferences make it impossible for her to be instrumentally rational.

Even in this version, the money pump argument fails if we assume *preference-based instrumental rationality*. First, in what sense do agents who are money pumped fail to maximize with regard to their own preferences? Note that the requirement invoked here sounds like *maximization* as we characterized it above. *Maximization* requires agents to choose such that no other option available to them is preferred to the one they choose. But note that the money pumped agents in fact never violate this requirement. In fact, we had to assume *maximization* in each individual choice to even get the result that the agent is money pumped. For each choice between a pair of options, the agent chooses her preferred one.

There must be another sense in which money pumped agents fail to maximize, if the argument is supposed to go through. It must be that they make a series of choices which leads them to an outcome that is not maximal amongst the outcomes they could have achieved had they performed a different series of choices. Let us call the requirement to not make such a series of choices *diachronic maximization*.[7]

[7] Synchronic and diachronic *maximization* are equivalent to what Gustafsson (2016) calls synchronic and diachronic dominance.

Money pumped agents only violate *diachronic maximization*, not our original synchronic *maximization*.

Gustafsson (2016) notes that diachronic maximization may not be as uncontroversial as synchronic maximization. However, he shows that we can give an agent with cyclical preferences a single choice that will force her to violate synchronic maximization. Namely, we can confront her with a single choice between all the outcomes amongst which she has cyclical preferences. In the apartment case, I would be offered the choice between all three apartments straight away. And in the Self-Torturer Problem, you would be asked to choose between all the different settings in one single choice.

Given the agents in these examples have cyclical preferences, no matter what they choose, they will end up with an outcome to which another available outcome would have been preferred. Thus, they are bound to violate synchronic *maximization*. Note that by similar reasoning, agents with cyclical preferences violate diachronic maximization even if they manage to refuse to trade early in the sequence of proposed trades. No matter which outcome they end up with, they could have ended up with a preferred one by stopping the trading earlier or later.

Now the question is what this shows. Gustafsson, for one, calls the single choice decision problems just presented 'synchronic money pumps' (even though no pumping seems to be going on), and thinks that the violation of *maximization* they entail for agents with cyclical preferences shows these preferences to be irrational. In fact, the argument that cyclical preferences are irrational because they make it impossible for agents to follow *maximization* can be found in earlier literature. Levi (2002) makes the same argument, and so do Davidson et al. (1955), just before they present their money pump argument:

> A rational choice (relative to a given set of alternatives and preferences) is one which selects the alternative which is preferred to all other alternatives; if there are several equivalent alternatives to which none is preferred, then any of these is selected. In short, a rational choice is one which selects an alternative to which none is preferred. But it is clear that the set of [cyclical] preferences makes a rational choice impossible, for whichever alternative [the agent] chooses there will be another alternative which is preferred to it. (p. 145)

However, I think that this is the wrong conclusion to draw from the money pump arguments and the observation that agents with cyclical preferences may fail to maximize. Instead, we should conclude that *maximization* is not a plausible principle of instrumental rationality for agents with cyclical preferences.

Preference-based instrumental rationality in fact only provides an obvious justification for *maximization* for agents who already have preferences that form a weak ordering over outcomes. By maximizing with regard to their preferences, these agents choose an outcome that is weakly preferred to all other outcomes. They thus seem to do best given their preferences: None of their binary preferences are frustrated. But this justification of *maximization* only works for agents whose preferences can be maximized.

If in our justification for *maximization* we already took for granted *weak ordering*, then we are not permitted to appeal to *maximization* when justifying *weak ordering*.[8] We would be justifying each of the two central requirements of standard decision theory with reference to the other. Instead, we are only permitted to treat *maximization* as a principle of choice appropriate for agents whose preferences form a weak ordering of outcomes, not as one that is necessarily appropriate for agents with cyclical preferences.

For agents who have cyclical preferences, as defenders of acyclicity point out, it may be the case that no outcome is such that it is weakly preferred to all other available ones. But this just means that

[8] For similar reasons, Andreou (2016) claims that the argument under discussion here is even question-begging.

maximization is not a plausible principle of instrumental rationality for agents with cyclical preferences. *Maximization* is a good principle of choice for agents with transitive and complete preferences because it ensures the agent does best by her preferences. A different principle of choice might ensure that an agent with cyclical preferences does best by her preferences.

What Gustafsson, Levi and Davidson et al. seem to ignore is that *maximization* is itself a principle that needs to be instrumentally justified, if decision theory is supposed to be a theory of instrumental rationality. And so it needs to be shown that *maximization* serves the agent's relevant conative attitudes as they are. Under *preference-based instrumental rationality*, the relevant conative attitudes are her preferences over outcomes. *Maximization* can be justified if those preferences are transitive and complete. But it does not qualify as a good principle of choice for agents with cyclical preferences, because it is impossible for those agents to maximize. But that need not mean that they cannot choose an action that serves their preferences as well as possible.

In fact, several choice rules for intransitive preferences have been proposed. For instance, according to a rule proposed in Schwartz (1972), an agent should choose a member of a subset of the available outcomes such that (1) no outcome outside of the subset is strictly preferred to any member of the subset, and (2) no proper subset of this subset fulfils condition (1). In our examples, if the agent is given a single choice between the outcomes over which she has cyclical preferences, according to this rule she is permitted to choose any of the options. At the same time, in each of the binary choices, the agent is required to choose the outcome she prefers.

Schwartz's rule is still in the spirit of *preference-based instrumental rationality.* It appears to be a good candidate for a rule that captures what it means for an agent with cyclical preferences to do well given her preferences. Such an agent will have to frustrate some of her preferences. But the rule identifies a set of outcomes that seems to ensure no preferences are frustrated unnecessarily. For instance, suppose there is a fourth Apartment D, which is tiny, windowless, and has a 2-hour commute. If I prefer all of Apartments A, B, and C to it, then Schwartz's rule would not let me pick this apartment. In the following, I will treat Schwartz's rule as our candidate choice rule that expresses preference-guidance for intransitive preferences in a similar way as *maximization* expresses preference-guidance for transitive preferences.

Given the availability of plausible choice rules for intransitive preferences, the argument that cyclical preferences make *maximization*, and hence rational choice impossible does not go through. And so neither type of instrumental argument based on money pumps goes through under the assumption of *preference-based instrumental rationality.* Does that mean that money pump arguments are unsuccessful?

To answer this question, note that appealing to choice rules for cyclical preferences does not do away with the possibility of being money pumped, unless those choice rules keep the agent from being money pumped. Schwartz's rule, for one, seems to still lead to the agent being money pumped. In each individual binary choice, the rule requires the agent to choose the strictly preferred outcome. Nevertheless, something does indeed seem to go wrong if an agent is money pumped: An agent ends up paying to get something that she could have had for free. And so while we have shown that the arguments in favour of acyclicity based on money pumps that are typically given fail, we have not done away with the intuitive irrationality of being money pumped. The root problem, I want to argue, is the assumption of *preference-based instrumental rationality.* If it does not allow us to say what is instrumentally irrational about being money pumped, then so much the worse for *preference-based instrumental rationality.*

3.5 The True Cost of Being Money Pumped

Andreou (2016) notes that if we wanted to make the argument that Gustafsson and others want to make, the original money pump argument is just 'a distraction'. In order to see that an agent with cyclical preferences will violate *maximization*, one need not construct a story where an agent ends up paying for something she could have had for free. We can show that she violates synchronic *maximization* simply by confronting her with a single choice over all the outcomes. And we can show that she violates diachronic *maximization* by offering her trades between the outcomes without asking for money in return: In the first example, by the time I end up with Apartment C, I will have chosen an outcome over which I would prefer Apartment A, that is, an outcome I could have had.

However, intuitively, the money pump argument is more than a distraction. What makes the series of choices we described seem instrumentally irrational is that the agent will end up where she started, but having lost money. If we think that only being money pumped repeatedly is irrational, that is because the agent ends up where she started, having lost significant amounts of money. If the agent were to end up with any of the apartments without having lost that money, we do not intuitively think that something is instrumentally amiss with her.[9] And in fact, as we will see in the next chapter, several authors have responded to money pump arguments by proposing that agents with cyclical preferences need not be money pumped if they have foresight. If they know that they will be offered a series of trades, they will refuse to trade early on. These authors assume that agents who manage to stop in this way avoid the pragmatic disadvantages the money pump argument tried to illustrate. And intuitively they do. Hence, arguments like Gustafsson's do not seem to capture the significance of the money pump arguments.

What is at fault here is the assumption of *preference-based instrumental rationality.* We have already seen that *preference-based instrumental rationality* does not let us distinguish between frustrating our preference between Apartment A and Apartment A - ϵ, and frustrating our preference between Apartment A and Apartment B. If preferences between outcomes are the basic standard of instrumental rationality, we cannot say that frustrating the first is instrumentally irrational and frustrating the second need not be. But money pump arguments rely on the intuition that this is so.

Note that Schwartz's criterion, too, cannot distinguish between the two preference loops just described, precisely because it is a criterion that only appeals to preferences between outcomes. If it is permissible to choose all of A, B, and C when choosing amongst A, B and C, then it is permissible to choose A, B, C and A - ϵ when choosing between those four. So suppose the rental agency lets you choose between Apartments A, B, and C, and leaves it up to you whether to pay an optional \$25 fee for choosing Apartment A. According to Schwartz's criterion, you would be rationally permitted to pay the \$25. But that does not seem right. Schwartz's criterion does not even condemn being 'synchronically money pumped' as irrational.

Moreover, no choice rule for intransitive preferences that only appeals to preferences between outcomes could have the result of allowing the agent to pick Apartment C but not allowing her to pick Apartment A - ϵ. These two options are treated symmetrically by the agent's preferences. Each is preferred to one other outcome, and dispreferred to two. In order for us to formulate a choice rule for cyclical preferences that distinguishes between making a choice between the options in the choice loop

[9] That is, unless the wasted time and inconvenience involved in switching are themselves significant costs. If there are such costs, the danger is not only that the agent is money pumped, but also that she is 'time pumped' or 'convenience pumped'. And then the point is that we wouldn't think there is anything wrong with ending up with any of the outcomes without having payed any of these costs.

not involving A - ϵ, and those in the choice loop involving A - ϵ, we need to privilege the preference between A and A $-\epsilon$ as one that may not be frustrated. In order to do so, it seems like we need to abandon *preference-based instrumental rationality.*[10]

We have seen in the last section that arguments for the irrationality of cyclical preferences do not work under the assumption of *preference-based instrumental rationality.* I have argued here that the assumption of *preference-based instrumental rationality* in fact keeps us from explaining what is intuitively irrational about being money pumped. Moreover, the assumption makes it difficult to formulate choice rules for cyclical preferences that may prevent an agent from being money pumped. As we have seen, choice rules that only take into account preferences between outcomes cannot do the job.

Seeing that money pumps have such an intuitive hold on us, we thus have good reason to give up *preference-based instrumental rationality.* In the next section, I want to argue that abandoning *preference-based instrumental rationality* makes independent sense. The next chapter argues that this abandonment does have far-reaching implications, however. Money pump arguments in favour of acyclicity ultimately only go through for agents who already have a specific kind of desire.

3.6 Desire-Based Instrumental Rationality

Preference-based instrumental rationality treats as basic preferences agents have between the objects of choice. In the case of choice under certainty, these objects of choice are outcomes. Preferences over outcomes form the standard against which actions or choice rules are judged. We already noted in the last chapter that this notion of instrumental rationality differs from the traditional picture whereby instrumental rationality is about desires. The main difference, I here want to argue, lies in the object of the relevant conative attitude.

The only preferences most orthodox decision theories deal with in the context of certainty are preferences over outcomes.[11] Even if we admit other preferences, preferences over outcomes are privileged in that they form the preferences over what we take to be the object of choice. As we said, outcomes are descriptions of all the circumstances an action may lead to that the agent may care about. In our first example, at a minimum, the outcome whereby I choose Apartment A would consist in a description of the size of the apartment, the length of my daily commute, and the views from the apartment. In the self-torturer case, the outcomes would include descriptions of both the level of pain I will be in, and the amount of money I will have.

One might think that appealing to 'what the agent may care about' here already presupposes a notion of caring about something that is independent from preferences. Yet, decision theorists have implemented this idea in a way that again appeals to preferences over outcomes. Joyce (1999, p.52) cashes out the rule for specifying outcomes as follows: Whenever there is some circumstance such that an agent would strictly prefer an outcome in the presence of that circumstance to the same outcome in the absence of that circumstance, the outcome has been underspecified.

Clearly, this rule for specifying outcomes will lead to outcomes being very detailed descriptions of states of affairs that may come about as the result of my actions. In fact, moving to more detailed

[10] In the Self-Torturer Problem, Schwartz's criterion allows the agent to pick any setting. As Carlson (1996) notes, this is already counterintuitive, even without appealing to the outcome where the agent has been money pumped. The agent who ends up rich, but in excruciating pain seems instrumentally criticizable. But again, it is difficult to say why without abandoning *preference-based instrumental rationality.*

[11] This is the case, in particular, for Savage (1954) and von Neumann and Morgenstern (1944).

descriptions of outcomes is a common move made in order to accommodate apparent violations of the standard axioms of decision theory, including transitivity.[12] And so those who want to defend transitivity need to embrace very fine-grained outcomes as the object of choice. But it is the fine-grained nature of these outcomes that makes them implausible candidates for the object of the conative attitude that should form the standard of instrumental rationality.

What I here want to highlight is that we do not ordinarily think of such detailed descriptions of states of affairs as the objects of our desires, even if they are the object of choice under certainty. What we claim to desire in ordinary discourse are simpler states of affairs. For instance, I might say, 'I desire to drink a glass of fizzy water right now', 'I desire to have more time to practise viola', or 'I desire to sail the Inside Passage'. In the last case, the object of my desire is not a complete description of one course that my life could take in which I sail the Inside Passage - complete with the description of what flavour ice-cream I will have after dinner tonight. The object of my desire is simply my sailing the Inside Passage. As Heath (2008) puts it, our desires are not usually 'all-things-considered' judgements over complete states of affairs (p.23).

We also sometimes speak of preferences with regard to such simpler states of affairs. I may say that I prefer sailing the Inside Passage to sailing the Northwest Passage, for instance. The objects of this preference are not fully specified outcomes. In fact, there are many outcomes involving me sailing the Northwest Passage that I would prefer to many outcomes involving me sailing the Inside Passage. I will prefer an outcome involving me sailing the Northwest Passage if that outcome also involves you giving me a million dollars by the end of it. But this does not make it any less true that I prefer sailing the Inside Passage to sailing the Northwest Passage in the ordinary sense just described. My desire to also have a million dollars is irrelevant to this preference.

While some decision theories, notably a Jeffrey-style decision theory, may feature preferences over these simpler states of affairs, these preferences are not directly about the objects of choice in decision-making under certainty. Those objects are fully described outcomes. We have said that *preference-based instrumental rationality* takes preferences over those outcomes to be the standard of instrumental rationality when evaluating an agent's choice. All preference-based arguments we have encountered so far, in fact, have done so.

The alternative, desire-based notion of instrumental rationality I want to propose here takes attitudes to simple states of affairs, like having another cup of coffee, to be the fundamental standard of instrumental rationality. I call it desire-based, since our desires ordinarily attach to such simple states of affairs. Desire-based instrumental rationality is meant to include, however, preferences between simpler states of affairs in the way we just described them. In fact, I take the statement that I prefer sailing the Inside Passage to sailing the Northwest Passage to be equivalent to the claim that I desire the former

[12] There is a debate on whether this move can be made ad infinitum, and whether this means that standard decision theory is normatively impotent, or is in need of non-Humean rationality principles. See, for instance, Broome (1991), Velleman (1993/2000) and Dreier (1996). As we see below, Pettit (1991) comes to a similar conclusion as I do here — namely that we don't need non-Humean principles of rationality, but rather need to rethink what conative attitude forms the standard of instrumental rationality. In fact, if Pettit is right, and what I call a 'desire-based' notion of instrumental rationality can help us solve the problem of how finely outcomes ought to be individuated, then this is another reason in favour of this notion of instrumental rationality. But note that the problems I point to here can already arise when outcomes are descriptions of more than one of the circumstances the agent cares about.

over the latter.[13]

One reason to err towards such a 'desire-based' notion of instrumental rationality is that it intuitively seems like desires for (or preferences over) simpler states of affairs are more basic, and explain preferences we may have over outcomes. If I prefer Apartment A over Apartment C, it is because I desire to have a beautiful view and a short commute, and this outweighs my desire to have a large living space. Outcomes, as full descriptions of everything the agent may care about, comprise many simpler states of affairs. Or, as we will sometimes put it, these simpler states of affairs are features of the outcome. Desires regarding states of affairs comprised in the outcomes are then potentially explanatory of the preference between the full outcomes. Pettit (1991) goes so far as to refer to a similar claim as a 'platitude of desiderative structure':

> There are two quite different sorts of object which desires may have - prospects and properties - and [...] the desires that we form for different prospects are determined by the properties that we think they have. Any prospects that we desire, any prospects that we prefer to the relevant alternatives, we desire for the properties they display or promise to display. (p. 151)

In the context of certainty, we may think of prospects here as outcomes. Pettit claims that we have more basic desires for outcomes to have certain properties - such as the property of involving me eating peanut butter and chocolate ice-cream. We desire outcomes for the properties they have, or, in our way of putting it, for the states of affairs they comprise.

Pettit thinks his 'platitude of desiderative structure' better captures our folk psychology. Another reason for thinking so is that it helps us better understand the phenomenon of having conflicting desires. I may sometimes prefer one outcome over another, because all my desires speak in favour of it. But sometimes, my preference will be formed in the context of many conflicting considerations. This is so, for instance, in the apartment case we have been considering. *Preference-based instrumental rationality* cannot make a distinction between the two cases and thus cannot account for the phenomenon of conflicting desires.

In fact, reflecting on the way we make decisions in the context of conflicting desires gives further credence to the idea that Pettit's 'platitude of desiderative structure' better captures our folk psychology. In these contexts, we do not generally come readily equipped with preferences over full outcomes. Trying to decide on where to live when moving to a new city can be very hard, even if I know all the relevant facts about the various options, as we are assuming in the context of certainty. Instead of consulting a preference over the relevant outcomes directly, I consult all the relevant desires I have regarding those outcomes. Any preference I form will be the result of weighing many conflicting desires. I may have to reflect long and hard about them to come up with a preference.[14] This seems to support the claim that in our reasoning processes, at least, desires (or preferences) over simple states of affairs are more basic than preferences over outcomes. Moreover, that we take our preferences to be answerable to our desires seems to support the claim that these desires are also the more fundamental standard of instrumental rationality.

[13] Note that preferences over outcomes or uncertain prospects are also sometimes understood as simply dispositions to choose one over the other, or as hypothetical choice. In fact, we will adopt such an interpretation of preference later (in Section 4.7). Note that that understanding of preference is fundamentally different from a comparative form of desire. Moreover, we couldn't, in turn, interpret preferences between simpler states of affairs in this way, since my choice between outcomes involving me sailing either passage will always depend on further features of those outcomes.

[14] In fact, the next chapter will argue that my desires will often not determine a unique right answer to what preferences I ought to have over outcomes.

If our preferences over outcomes are answerable to our more fundamental desires for states of affairs, this also creates the possibility that our preferences misrepresent our desires. This seems to be a kind of error that is familiar, but that *preference-based instrumental rationality* cannot explain, providing further support for a desire-based notion of instrumental rationality. The kind of error in question occurs when we form preferences over outcomes, potentially leading to decisions, that do not do justice to our desires. I may come to think that my preferences at a particular point in time were mistaken even if my underlying desires have not changed, and I have not learned new information. I may simply think that I did not do full justice to all my desires, or failed to weight them correctly. And if my actions were based on those preferences, then I may take this to be an instrumental failure.

For instance, suppose that what I care about most in my food is that it is healthy. Based on that, I correctly prefer the salad option over the pasta option in my university's cafeteria on most days. But if I continue to have that preference on a day when the salad option is *Wurstsalat*, and act on it, it seems like I have failed instrumentally. Similarly, the last chapter suggested that we can only give an instrumental argument in favour of resisting temptation if we can interpret the preferences of the tempted agent as mistaken in this kind of way: When I am tempted to watch a second episode, my momentary preferences fail to capture that my underlying desires really support going back to work. In the literature on temptation, at least, the idea that the tempted agent's preferences are mistaken representations of what she really desires is fairly widespread. Desire-based instrumental rationality may allow us to cast acting on such preferences as a failure of instrumental rationality.

Pollock (2006) provides another reason to abandon *preference-based instrumental rationality* in favour of desire-based instrumental rationality. And that is that it is 'computationally impossible' for sophisticated cognitive agents like us to have a database of the complete set of binary preferences over all the different fine-grained outcomes we may have to evaluate. It is simpler for us to store a data base of more primitive values that we use to compute what to do on each occasion. For him, too, something other than preferences between outcomes is basic. In particular, this is something akin to what Pettit called desires for properties of outcomes, and we have called desires for simple states of affairs, namely 'feature likings'. Pollock gives the example that he likes eating Greek food.[15]

Lastly, abandoning *preference-based instrumental rationality* in favour of desire-based instrumental rationality can help us explain what seems to go wrong when an agent is money pumped. It is the fact that the objects of preference in *preference-based instrumental rationality* are outcomes that kept us from explaining what goes wrong when an agent is money pumped. A, A - ϵ and B are all outcomes the agent has preferences between, and that's all we can say according to *preference-based instrumental rationality.* To give rational significance to the fact that A - ϵ is the same as A, except that the agent has less of some good she desires, we need to make features of outcomes instrumentally relevant.

As Andreou (2016) points out, the money pumped agent, at least in examples like that of the apartment choice, violates this principle:

[15] In fact, there is research in empirical psychology showing that agents tend to make better and faster decisions, and become less frustrated in the presence of a large and complex set of options, when decision-making is 'attribute-based' rather than 'alternatives-based'. That is, they make better decisions and become less frustrated when they break different options down into their various different attributes, rather than if they start by making direct comparisons between options. For instance, Huffman and Kahn (1998) show that consumers are more satisfied with their choices when they are first asked to indicate their preferences regarding general attributes, and are then presented with options described in terms of these attributes. Ford et al. (1989) present evidence that time pressure leads to increased use of attributes-based decision-making. Fellows (2006) shows that agents who have suffered damage to their prefrontal cortex are likely to use alternatives-based decision-making when healthy subjects use attributes-based decision-making, and suggests that this may explain their typically poor decision-making.

> P: It is irrational to make a choice or series of choices that leads one to an alternative y which is such that y is identical to another alternative x except with respect to one dimension of concern and, in that respect, y is dispreferred to x. (p. 1454)

According to this principle, an agent with cyclical preferences is irrational when she ends up with Apartment A having payed $25, be it in the single choice, or after a series of trades. But principle P makes reference to 'dimensions of concern', where concern here must be understood as a conative attitude other than preference over outcomes. In classical money pumps, the relevant dimension of concern is money — The agent has a desire for money. But as we said, the relevant dimension of concern could equally be free time, avoiding embarrassment, or keeping my secrets.[16] P seems like a plausible principle of instrumental rationality, because if it is violated, the agent has unnecessarily frustrated one of her desires. The money pumped agent is deprived of something she desires, but gets nothing else she desires in return.

Principle P can thus not be justified with reference to *preference-based instrumental rationality*, but only with reference to a desire-based notion of rationality. According to this notion, instrumental rationality requires an agent to do well by her desires, where the objects of desires are not complete outcomes, but simpler states of affairs.

What else can we say about the requirements of desire-based instrumental rationality? Desire-based instrumental rationality claims that an agent ought to do well by her desires over simple state of affairs. Since in the context of certainty, agents choose between full outcomes, the agent ought to do well by all of her desires over the states of affairs these outcomes comprise. As we said, conflicting desires will require the agent to somehow weight her various desires, and this may be hard. Still, we assume that some outcomes are clearly unacceptable in the light of the agent's desires, while others would be supported by the agent's desires.

Ending up money pumped, time pumped, dignity pumped, or secret pumped clearly seems to be an unacceptable outcome in the light of the agent's desires. In the following and in the next chapter, we will consider whether we can use that observation to argue that it is also a requirement of desire-based instrumental rationality that agents form acyclical preferences over outcomes in the light of their desires, and then maximize with regard to them.

3.7 Desire-based Instrumental Rationality and Instrumental Arguments

We saw in the last chapter that abandoning *preference-based instrumental rationality* has the advantage that we can give fairly straightforward arguments why resisting temptation may be instrumentally rational. If preferences are not themselves the ultimate standard of instrumental rationality, then they may misrepresent the true standard of instrumental rationality. Thus, if a temptation case, which we defined as a case of temporarily shifted preference, can be analyzed as a case where the shifted preferences misrepresent the agent's true underlying desires, then it will be instrumentally rational to resist temptation for straightforward reasons: It will serve the agent's actual desires.

Moreover, with something like principle P in hand, can we finally offer a successful money pump argument in favour of the acyclicity of preference? This is certainly what Andreou (2016) intended by

[16] In fact, those kinds of desires are better candidates for plausible money pumps, since they are not as clearly merely instrumental to something else, like money is.

introducing the principle. Abandoning *preference-based instrumental rationality* in favour of desire-based instrumental rationality does solve the problems we noted with earlier money pump arguments, since we now have a standard other than preferences over outcomes themselves to assess the agent's preferences over outcomes.

First note, however, that once we accept something like principle P, the "synchronic money pump argument", at least, does not work anymore. We said above that Schwartz's rule, which seems to capture what it means to be guided by one's preferences given their intransitivity, does not rule out choosing Apartment A - ϵ in a synchronic choice. However, adopting desire-based instrumental rationality, we can simply appeal to something like principle P in addition to Schwartz's rule. And then we can say that it would be irrational for me to choose Apartment A - ϵ, while it need not be irrational to choose either Apartments A, B, and C. In order to avoid being money pumped, I need not abandon my cyclical preferences.

So our only hope for a money pump argument to go through is in fact the diachronic money pump argument. An agent who is diachronically money pumped violates the diachronic part of principle P: She will engage in a series of choices that leads her to an alternative that differs from another one she could have had only by having less of some good she desires. If the agent is guided by her preferences in her choices, by applying Schwartz's rule, then she will be money pumped.

In this case, therefore, we cannot simply add principle P as a further condition on top of Schwartz's rule in order to rule out that the agent is money pumped. If the agent is to be guided by her preferences in action, then she must change her preferences in order to avoid being money pumped. Proponents of standard decision theory would like to argue that the agent should make her preferences acyclical and maximize with regard to them in order to avoid being money pumped, and thus abide by principle P.

Any such argument relies on the validity of the diachronic aspect of principle P, namely that it is irrational to make a series of choices that leaves you money pumped. It is controversial whether there are such diachronic principles. Hedden (2015b) argues that what he calls susceptibility to 'diachronic tragedy' alone does not necessarily make an agent irrational. The diachronic part of P may appear too strong a requirement especially in cases where an agent's desires change over time. As we saw in the last chapter, in such cases, avoiding diachronic tragedy may require taking an action that is not supported by the desires the agent has at the time of action regarding the outcomes still available to her then. In those kinds of cases, insisting on the diachronic part of principle P seems to be in conflict with insisting on the instrumental rationality of the individual actions.

To sidestep these worries, we can formulate a weaker version of the diachronic part of principle P. I will call it principle Q:

Q: It is irrational to make a series of choices when

1. it leaves one with an outcome y, when there would have been an alternative series of choices which leaves one with outcome x, which is identical to y, except that it has more of something one desires,[17] and
2. there is an alternative series of choices of which (1) is not true, and whereby each individual choice would have been supported by one's desires concerning the outcomes still available at the time of action.

[17] I mean these objects of desire to be features of outcomes that are individuated finely enough such that they themselves are not desired by the agent in one respect, while she is averse to them in another. For instance, if there is some respect in which I am averse to having more money, I do not violate principle Q when I am money-pumped.

This principle claims that it would be irrational for an agent to be money pumped if she could have avoided being money pumped by taking a series of actions each of which is itself unproblematic in terms of desire-based instrumental rationality. The sense of "money pumped" I have in mind here is wide in scope. I take an agent to be money pumped in this wide sense whenever she ends up with an outcome that differs from another outcome she could have had merely by lacking, or having less of something she desires.

If money pump arguments are to be successful, then we must hold that money pumped agents also violate this weaker principle. The acyclical preferences that these arguments want to show the agent should have will presumably endorse a course of action whereby the agent avoids being money pumped. If there is no course of action that does not lead to being money pumped in which each action is itself endorsed by desire-based instrumental rationality, then no acyclical preferences seem to exist that endorse a course of action that is itself instrumentally unproblematic. And then we will have failed to give an instrumental justification for acyclical preferences.

This principle thus finally seems to give us a basis for making a money pump argument. Note also that it allows us to deal with the special case of temptation we appealed to in the last chapter. Once we abandon *preference-based instrumental rationality*, we will have to say more about how preferences ought to relate to the true standard of instrumental rationality, namely desires. The next chapter will say more about how we may understand preferences in the light of the agent's various desires. For now, let us assume that the agent's preferences over outcomes are responsive to her desires, and the result of some kind of weighting activity between the agent's various desires.

It is then possible that several different preference orders over outcomes represent an agent's underlying desires well. The next chapter will argue that this is in fact likely in the kinds of problem cases for standard decision theory we have been considering. Given this possibility, it might be that in some cases of preference shift, there is a preference relation that would have been an accurate reflection of the agent's desires throughout.

Appeal to principle Q now can form the basis of an instrumentalist argument that the agent should either stick to those constant preferences throughout the episode of temptation, and act in accordance with them, or to act as if she did. This is because we can interpret the agent who ends up with the 'tragic' outcome of not watching any TV in our example (see Section 2.4) as being money pumped in the wider sense of the word.

Presumably our agent desires both getting work done and watching TV. Suppose that preferences that rank watching one episode most highly represent the agent's desires well at any point in time, but that the shifting preferences we originally described also do. If she sticks with those shifting preferences, acts in accordance with them, and does not watch any TV, she ends up with an outcome that differs from watching one episode only by having less of something she values: She watches no TV, but also does not get any more work done than she would have had she watched TV. That is, she payed a cost of pre-commitment (not watching any TV) for something she could have had for free (getting all her work done). At the same time, watching one episode of TV would not have meant acting in a way that is not itself endorsed by desire-based instrumental rationality. This is because the agent's desires at the time of temptation also would have supported a preference ranking of the outcomes that ranks watching one episode most highly. Principle Q thus endorses acting in accordance with the constant preferences in this special case.

This argument is two-tier in the sense that it makes us look at the instrumental advantages of an entire sequence of actions rather than merely at individual actions. For this one special case, then, we have thus identified an instrumentalist two-tier argument for resisting temptation, and moreover one that does not rely on claiming that the agent's preferences under temptation are a false representation of her underlying desires. If money pump arguments are to be successful, they offer a two-tier account in the same sense. This is because principle Q, which they are both based on, is a principle about the instrumental benefits of an entire series of choices.

We may, however, still be worried that principle Q is too strong. We may worry that it condemns as irrational agents we may not feel comfortable calling irrational. Take, for instance, the example of Abraham and Isaac as presented by Broome (2001). God asks Abraham to sacrifice his son. Abraham is deeply torn between being obedient to God, and saving his son. Indeed, he finds these two values incommensurable. Suppose that under conditions of incommensurability, either choice is permissible. Suppose, then, Abraham sets out to the top of the mountain in order to sacrifice his son. Half-way up, he reconsiders his choice. Trust between father and son is already damaged, but even so, both choices still seem incommensurable. If Abraham then goes back to the foot of the mountain, he will end up with an outcome he could have had without sacrificing the trust of his son. He thus seems to have violated principle Q.

One might worry here that, while there has been this loss, Abraham has not done anything irrational. At every point in time, he chose an action that is permissible if we simply look at the way in which he evaluates the outcomes he can still achieve through his actions from then on. And then violating principle Q would not necessarily be irrational. As we will see, similar worries may arise for any way of avoiding being money pumped. Money pump scenarios turn out to be structurally similar to the case of Abraham.

A first way to respond here is simply to insist that some actions are irrational only in light of the series of actions they are a part of. This is a position that has gained popularity in the literature on diachronic rationality.[18] We may also note that the intuition that Abraham is not irrational seems to be driven by the trust that is lost between father and son to be a comparatively low cost, that is incurred only once. If Abraham were to keep on going up and down the hill indefinitely, perhaps even long enough such that the threshold of incommensurability is crossed, this seems more decisively irrational. This may convince some that it is already irrational when the cost is incurred only once. If not, Q could be reformulated in a way that only applies to being money pumped repeatedly.

Lastly, for the purposes of money pump arguments, it will in fact be enough to, for now, treat principle Q as a principle that is defeasible. That way, we can first look at the ways an agent may avoid being money pumped, and then consider whether failing to act in those ways is really irrational. It will turn out that money pump arguments only count in favour of acyclicity for agents who have a desire to have stable preferences over time. And for those agents, I will argue, the worries just described do not arise.

We just responded to the worry that principle Q is too strong due to its diachronic nature. However, we may also be worried that it is too weak. There are some series of choices that Q does not condemn as irrational, but that intuitively seem irrational for similar reasons as being money pumped in the apartment case and the special kind of temptation case just described.

First, as it stands, Q does not condemn being money pumped in the self-torturer case as irrational,

[18] See, for instance, Andreou (2006). Tenenbaum (2016) argues that principles of instrumental rationality also apply to extended actions. If we think of a series of choices as an extended action, then this could help us justify something like Q.

at least if we take adjacent settings to be in fact indistinguishable in terms of pain. If the relevant desires in this case are a desire not to feel pain, and a desire for money, then whatever setting the agent may end up with, it will be true that another setting would have been the same except for having more of something she desires. That will always be true of the next higher setting: It is the same in terms of the pain the agent feels, but would give the agent more money. And so even if the agent is money pumped, condition (2) in Q will not be satisfied.

Similarly, Q cannot deal with the following variation on the special temptation case: I do actually get a little bit more work done by not watching TV, than I would by watching one episode. It is just that this is definitely worse than watching one episode according to my desires at any point in time, and thus still 'tragic' in a wider sense of the word. While there is a conflict between my desire to watch TV and my desire to work, my desire to watch TV clearly outweighs my working desire in this circumstance.

Can we formulate a stronger principle of desire-based instrumental rationality that would condemn ending up with the 'tragic' outcome in both of these cases? In both cases we just considered, the agent's desires seem to allow us to rule out some outcomes as certainly unacceptable. We could use this observation to formulate a stronger, but also vaguer version of Q. In the TV case, even though the agent is not strictly money pumped when she does not watch any TV, it is also clear that the agent's desires clearly support sacrificing a little bit of work for watching one episode of TV — while at the same time the conflict between the desire to watch TV and the desire to work is not so clearly resolved when it comes to watching a second episode.

In the self-torturer case, the agent can also rule out some outcomes as clearly unacceptable. For instance, the outcome of having been money pumped several times, and thus losing significant sums of money will clearly be an unacceptable outcome to end up with according to the agent's desires. But so will many other outcomes in this case. Ending up at the last setting should also be a bad way of resolving the conflict between a desire for money and a desire to be pain-free for most agents. At the same time, not every setting will be clearly unacceptable in this way. After all, the agent's desires should permit her to choose *some* setting if she were to choose amongst them in a one-off choice.

Accordingly, I want to propose the following weaker, but also vaguer version of Q, and call it R:

R: It is irrational to make a series of choices when

1. it leaves one with an outcome which, according to the agent's desires at all points in time in the decision problem is unacceptable amongst the outcomes that could have been achieved in the decision problem.

2. there is an alternative series of choices of which (1) is not true, and whereby each individual choice would have been supported by one's desires concerning the outcomes still available at the time of action.[19]

[19] Andreou (2015) argues that we need to appeal to appraisal categories such as 'terrible', 'great' or 'fantastic' in order to express what is irrational about ending up at the last setting in the self-torturer problem. She formulates a principle that claims it is irrational to end up with an outcome in a (determinately) lower appraisal category than another outcome one could have had (p.570). That is, it is irrational to end up with a terrible outcome when one could have ended up with a great outcome. My appeal to 'unacceptable' can be understood as appeal to such an appraisal category. Moreover, any appeal to appraisal categories in a principle of instrumental rationality implies one has given up on the idea that preferences over outcomes form the standard of instrumental rationality. My claim is that an appraisal category like 'unacceptable' can be made sense of within a desire-based notion of instrumental rationality. Since desire-based instrumental rationality also provides the rationale for principles P and Q, it has the resources to unify the treatment of money pump arguments and the self-torturer problem that I have argued, in agreement with Andreou, to be most promising.

We cannot say anything more concrete about what would make an outcome an unacceptable one to end up with in a decision problem without going into detail on how conflicts between desires ought to be resolved into a choice. One clear case in which an outcome is unacceptable according to the agent's desires is when an agent is money pumped while she could have avoided being money pumped. In that case, one of the agent's desires is frustrated unnecessarily. Q is thus implied by R. Where we can, we will thus stick to principle Q, which is on much firmer ground. The next chapter will consider whether appeal to such a principle can finally help us make a money pump argument in favour of the acyclicity of preference.

3.8 Conclusions

Since standard decision theory is formulated in terms of preferences, it may seem natural to adopt *preference-based instrumental rationality* when defending it as a theory of instrumental rationality. This chapter argued that, surprisingly, *preference-based instrumental rationality* in fact makes it impossible to give an instrumental justification of one of the central requirements of standard decision theory in the context of certainty, namely the acyclicity of preference. *Preference-based instrumental rationality* makes it impossible to make a money pump argument in favour of acyclicity.

I argued that the basic flaw of *preference-based instrumental rationality* is that it makes preferences over outcomes the standard of instrumental rationality, where outcomes are detailed descriptions of complete states of affairs. But our basic desires and preferences typically have simpler states of affairs as their objects. Instrumental rationality, I argued, should be about doing well by those basic desires. It is these basic desires we need to appeal to in order to explain what may be wrong with an agent who is money pumped (or time pumped, or dignity pumped).

By motivating principle Q, this alternative notion of instrumental rationality gives us the basis for formulating a money pump argument in favour of the acyclicity of preference after all. Desire-based instrumental rationality also makes it possible to give an instrumental argument in favour of resisting temptation. First, as we argued in the last section, if we give up *preference-based instrumental rationality*, it becomes possible that our preferences misrepresent the conative attitude that forms the true standard of instrumental rationality, namely more basic desires. If this is so in a case of temptation, the agent has a straightforward instrumental justification for resisting temptation.

But, as we have just seen, desire-based instrumental rationality also seems to make possible a two-tier argument for stability of preference in a special kind of case, one where multiple preference orders represent the agent's desires well. The next chapter argues that, in fact, such 'non-uniqueness' is to be expected, especially in problem cases for standard decision theory. However, I also show that this is deeply problematic for the money pump arguments.

The next chapter will turn to whether money pump arguments in favour of acyclicity really can be salvaged under the assumption of desire-based instrumental rationality, with principle Q in hand. I will argue that the new challenge for money pump arguments is that we can no longer take *maximization* for granted as a principle of rationality under the new desire-based instrumental rationality. The best argument we can make for acyclicity, I claim, involves redefining preference not as a conative attitude, but as a disposition to choose. And even then, the argument only applies to agents who have a specific kind of desire, one that favours stable dispositions to choose, to begin with.

Chapter 4

Desire-Based Instrumental Rationality and the Money Pump Argument

4.1 Introduction

If standard decision theory is to serve as an instrumental theory of rationality, we need to give instrumental justifications for its core requirements. In the context of certainty, these are *weak ordering* and *maximization.* Most decision theorists abide by what we called *preference-based instrumental rationality*, according to which preferences over outcomes form the basic standard of instrumental rationality. This notion of instrumental rationality makes *maximization* plausible as a principle of rationality, if the agent's preferences are already well-ordered.

However, the last chapter argued that *preference-based instrumental rationality* also makes it impossible to give an instrumental justification for acyclicity, which is strictly weaker than *weak ordering.* I instead advocated a desire-based notion of instrumental rationality. Here, I want to investigate whether this notion allows us to finally give instrumental justifications for both core requirements of standard decision theory under certainty.

It may certainly seem like appealing to the agent's underlying desires over simple states of affairs may help us make money pump arguments in favour of acyclicity. The last chapter argued that a desire-based notion of instrumental rationality can ground principle Q:

> Q: It is irrational to make a series of choices when
>
> 1. it leaves one with an outcome y, when there would have been an alternative series of choices which leaves one with outcome x, which is identical to y, except that it has more of something one desires, and
> 2. there is an alternative series of choices of which (1) is not true, and whereby each individual choice would have been supported by one's desires concerning the outcomes still available at the time of action.

According to this principle, it is irrational to be money pumped, that is, to forego something one desires unnecessarily, in circumstances where avoiding being money pumped does not conflict with the instrumental rationality of each individual action. Desire-based instrumental rationality thus at least enables us to explain why it is sometimes irrational to be money pumped.

However, this chapter argues that this is not enough for the money pump arguments to go through. Moreover, much of this has to do with the fact that under desire-based instrumental rationality, there is no straightforward justification for *maximization*, or preference-guidance more generally, anymore. And then the basic problem is this: Pointing out that it is irrational to be money pumped is insufficient for justifying acyclicity, if there are alternative ways of avoiding being money pumped than adopting acyclical preferences. Failing to always be guided by one's preferences in action is one such alternative way.

Abandoning a preference-based notion of instrumental rationality for a desire-based one raises the question of what preferences over outcomes are, and how they relate to desires. There seem to be two main answers to this question. On the first, preferences are still conative attitudes over outcomes, and act as a (fallible) summary of the agent's various desires over the simpler states of affairs these outcomes comprise. On the other, preferences are interpreted to be dispositions to choose. Accordingly, this chapter proceeds in two parts.

In the first I will argue that if preferences are conative attitudes that summarize the agent's desires, desire-based instrumental rationality can justify neither *maximization* nor *acyclicity.* In the second, I argue that if preferences are dispositions to choose, desire-based instrumental rationality can justify the two principles only for agents who have a desire to have stable choice dispositions over time, and across different choice situations and contexts.

I will thus conclude that the best case for the requirements of standard decision theory under certainty involves going with the second interpretation of preference as choice disposition. But even this best case only gives an instrumental justification for these requirements conditional on the agent having a specific kind of desire, namely the desire to have stable choice dispositions. And so standard decision theory, at best, only gives us conditional principles of instrumental rationality.

4.2 Foresight as an Alternative Response to Money Pumps

I have argued that it is instrumentally irrational to be money pumped according to desire-based instrumental rationality, at least as long as it is possible to avoid being money pumped without taking actions that are themselves suspect in terms of the agent's desires at the time of action. This will only give an agent conclusive reason to have acyclical preferences if having acyclical preferences is the only way of avoiding being money pumped.

However, it is frequently argued that having foresight can save an agent with cyclical preferences from being money pumped without her having to give up her cyclical preferences. Here I consider whether foresight can really save the agent from being money pumped. The short answer is that it can only do so if the agent violates *maximization* at some point. However, as we will see in the next section, such a violation need not be in conflict with desire-based instrumental rationality.

Schick (1986) argues that an agent with foresight will stop trading early in the series of trades offered to her in money pump scenarios. Let us consider his argument in the context of the example of my apartment choice introduced in the last chapter. First, we need to represent the series of choices I am offered by the rental agency as a dynamic choice problem. As we saw in Chapter 2, such problems are usually represented in decision trees like the one in Figure 4.1. At the square nodes, an agent can decide whether to go up or down, starting at the left-most node. We assume here that the agent is not offered further trades once she has refused one.

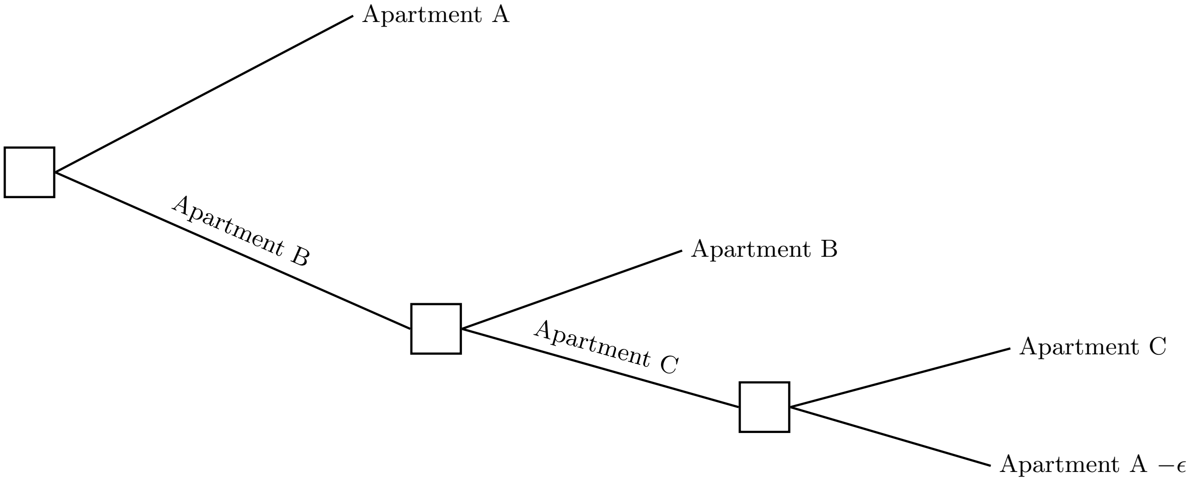


Figure 4.1: Dynamic Money Pump Problem

Schick argues that in a decision problem like this one, agents should decide using a process of backward induction. They should consider how they will choose in the last choice, assuming they will be guided by their preferences over the outcomes available then. They then take their prediction as given when considering the second to last choice, etc. As we noted in Chapter 2, this approach to dynamic choice has come to be known as 'sophisticated choice'. We there characterized it as continued application of *maximization*, coupled with the belief that one will maximize in the future.

Now that the agent has cyclical preferences, we cannot assume she follows *maximization*. But as we saw in the last chapter, we can still make sense of the idea of preference-guidance in the context of intransitive preferences, for instance by appealing to Schwartz's rule. In the binary choice the agent faces here, this rule, like *maximization*, implies that the agent should choose the option with the strictly preferred associated outcome. We can thus still make sense of sophistication when the agent's preferences are cyclical: It can be characterized as continued application of some rule to be guided by one's preferences in action.

A sophisticated agent facing the series of trades in our example will indeed not be money pumped, but instead end up with Apartment B. She will predict that, if she got to make the last choice, she would choose to pay $25 to get Apartment A back. To predict what she would do at the second to last choice, we thus consult the agent's preferences between Apartment B and Apartment A $-\epsilon$. Since Apartment B is preferred to Apartment A, it is also preferred to Apartment A having payed $25. And so the agent would choose to stick with Apartment B in the second to last choice. Knowing that, she would choose Apartment B in the first choice.

However, Rabinowicz (2000, 2001) shows that sophistication will not save an agent from being money pumped in all situations. Appeal to sophistication alone thus does not decisively speak against money pump arguments in favour of acyclicity. In particular, sophistication may not save an agent from being money pumped in cases where she faces an opponent that is persistent.

To modify Rabinowicz's (2000) example slightly, and adapt it to our purposes, suppose that the rental agency offers me a trade exactly three times, even if I have refused the first or second trade. I start with Apartment A. They offer me Apartment B. If I accept the offer, they offer me Apartment C. If I accept that offer, they offer me Apartment A - ϵ. If I have refused a trade, they simply offer me the same trade again. The dynamic choice problem I now face is represented in Figure 4.2.

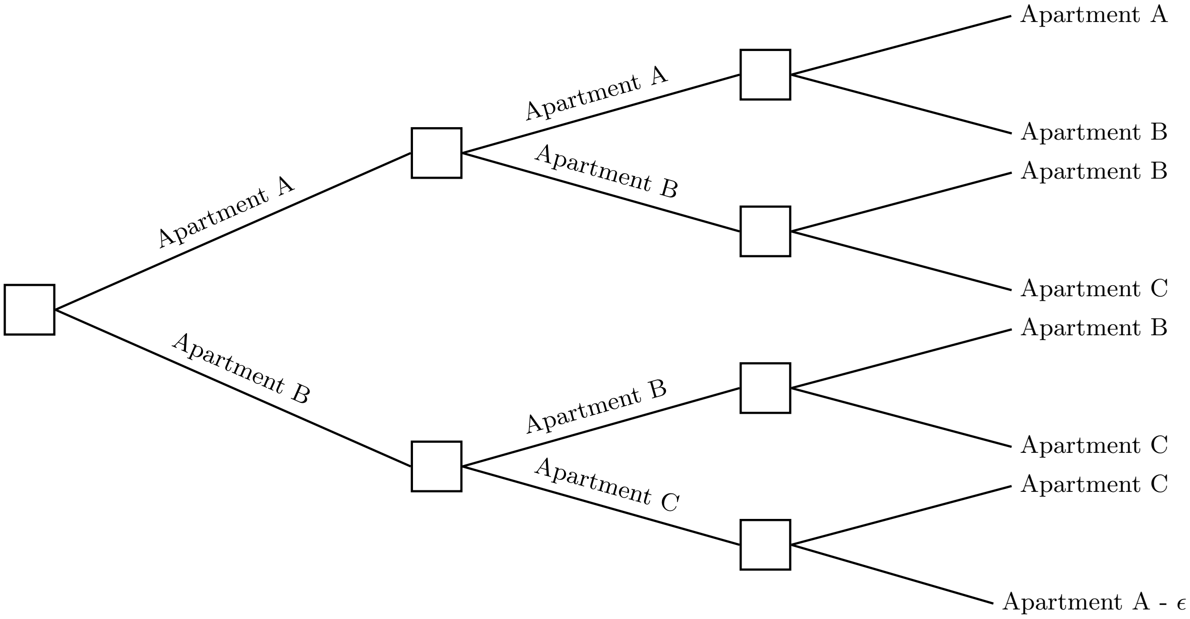


Figure 4.2: Modified Dynamic Money Pump Problem

In this modified money pump, even a sophisticated agent will be money pumped. That is, she will end up with Apartment A - ϵ even though she could have had Apartment A. To see that, note that at each of the last decision nodes, the sophisticated agent will choose to trade. Given that, she will also trade at the second to last choice node. The choice the sophisticated agent thus effectively faces at the first choice node is between Apartment C (if she refuses to trade now), and Apartment A - ϵ (if she trades straight away).

At the first choice node, the sophisticated agent knows she will trade in the future, and thus knows that Apartments A and B are inaccessible to her. Given a choice between Apartment C and Apartment A - ϵ, she chooses the latter. Thus, we have shown that sophistication does not save the agent from being money pumped. We only need some persistence on the side of our rental agency. Does this mean that foresight in general does not help?

It has been proposed that different ways of deciding in dynamic choice problems may help an agent with cyclical preferences to never be money pumped. One such choice strategy is what McClennen (1990) calls 'resolute' choice, as we already encountered it in Chapter 2. Another is what Rabinowicz (2014) refers to as 'unified choice'. Resolute agents choose a sequence of actions in the beginning of the dynamic choice problem, in accordance with their preferences then, and then simply go through with that sequence of actions. Unified agents decide at each point in a sequence of choices as they would were they to *then* decide on the whole sequence in one single choice. Given stable preferences over outcomes, this should lead to the same course of action. Since this stability will be presupposed in all dynamic choice problems from now on,[1] Rabinowicz' unified choice should be understood as included whenever I mention resolute choice.

In our case, a resolute agent would choose as she would in the 'synchronic' money pump. Rabinowicz (2014) claims that the agent would not choose to be money pumped in this case:

> Were she to make a single choice, [...] we may safely assume, she would not choose to accept all the three exchanges. A simple calculation would show that refusing all of them would save her the extra costs and still result in the same outcome. (p. 373)

Note here that this solution seems to rely on abandoning *preference-based instrumental rationality.* As we have seen, as far as *preference-based instrumental rationality* is concerned, we cannot say that the agent is rationally required to not choose A - ϵ. But money pumps only have an intuitive hold on us in the first place because we think it is in fact instrumentally irrational to end up paying for something you could have had for free (especially in a one-off choice). We can appeal to principle Q (Section 3.7) to explain why a resolute agent should not select a course of action that would leave her money pumped. And so neither money pump arguments, nor the most common response to them work without abandoning *preference-based instrumental rationality.*

Resolute, or unified choice, combined with something like principle Q, keeps an agent with cyclical preferences from being money pumped. Does this mean that the money pump arguments have been refuted? The central controversy here has been that it is not clear whether resolute choice is itself compatible with instrumental rationality. Moreover, if it is not, then it is not clear whether our principle Q really condemns being money pumped here as irrational.

This worry is usually expressed in terms of *preference-based instrumental rationality*, and thus, I think, misses the point. Sophisticated agents make a prediction of their own future behaviour, assuming they will be guided by their preferences over the outcomes still available at each choice point in the future. Given this prediction, they choose the action that is favoured by their preferences over the outcomes that are still available to them. Sophisticated agents thus merely seem to act continually in accordance with their preferences. If we think that being guided by one's preferences, by following *maximization* or Schwartz's rule, is a principle of instrumental rationality, then instrumental rationality seems to demand that an agent be sophisticated.

By contrast, on one natural way of understanding resolute choice, a resolute agent will sometimes choose against her binary preferences over the outcomes still available to her at the time of action. Namely, this is so if the resolute agent's preferences throughout remain the same as they would be outside of the dynamic choice problem. We already saw in Chapter 2 that resolution understood in this way may require counter-preferential choice in temptation cases. In the present case, the agent will act against her preferences at some point whenever resolute and sophisticated choice come apart.

For instance, in the modified money pump we just presented, resolution demands refusing at least one trade, since the resolute agent would not choose to end up with A - ϵ. If the resolute strategy is to refuse the last one, then the agent acts against her preferences over the available outcomes then. If it is to refuse the second to last one, while taking the last one, again, taking into account a correct prediction of one's future action, the agent then acts against her preferences over the available outcomes. The same

[1] That is, stability of the agent's 'given' preferences will be presupposed. Since on one way of understanding resolution, resolution involves making temporary adjustments to one's preferences, resolution may itself introduce instability in the agent's ultimate preferences. What we will call 'given' preferences are the preferences the agent has outside of specific dynamic choice problems, or before preferences have been adjusted as part of a resolute choice strategy.

holds if the first trade is the one to be refused. Thus, resolution sometimes demands acting against one's preferences in money pump scenarios.

On this understanding of resolution, the only way that foresight can keep an agent from being money pumped is thus by sometimes acting counter-preferentially. We have said that *maximization*, or at least preference-guidance in the form of Schwartz's rule seem to be easily justifiable under *preference-based instrumental rationality.* As we saw in Chapter 2, two-tier arguments in the face of temptation that try to argue otherwise fail. Being guided by one's preferences in action seems to guarantee that one does well in terms of one's preferences. And so, if we assume *preference-based instrumental rationality*, resolution does not speak against money pumps. This is because according to *preference-based instrumental rationality*, resolution is itself incompatible with instrumental rationality.

We come to a similar conclusion on another reading of resolute choice. McClennen in fact suggests that resolute agents adjust their preferences within dynamic choice problems, so that being guided by one's preferences leads one to the resolute choice after all. If, for instance, the resolute choice would require us to end up with Apartment C, the agent would adopt the preference $C \succ A - \epsilon$ at the last choice node in the dynamic choice problem. If agents adjust their preferences as part of a resolute choice strategy, resolute agents in fact end up maximizing at each point in time with regard to those altered preferences. What then still distinguishes it from sophistication? To preserve what I think is a meaningful difference, we have to slightly amend our characterization of sophistication: Sophistication in fact consists in the continued application of some preference-guidance norm like *maximization* to the agent's 'given' preferences. I will take the agent's 'given' preferences to be those preferences the agent would have outside of a particular dynamic choice problem, or before adjustments to preferences have been made as part of a resolute choice strategy.[2]

One could be worried that one cannot have instrumental reasons to make changes to one's preferences like this. And then being resolute would not be a strategy for avoiding being money pumped that one has instrumental reasons to adopt, but just a circumstance some agents are in, and others are not. Again, *preference-based instrumental rationality* supports these worries. If preferences themselves are the basic standard of instrumental rationality, being resolute in this alternative sense requires the agent to change what she ultimately cares about. And then it seems like any reason to be resolute would be a non-instrumental reason. One caveat here is that having different preferences can serve an agent's preferences as they are in cases of 'autonomous benefit'. However, I argued in Section 3.4 that money pumps cannot be construed as such cases of autonomous benefit under the assumption of *preference-based instrumental rationality.*

Hence, for both interpretations of resoluteness, standard worries rely on *preference-based instrumental rationality.* But we have argued in the last chapter that *preference-based instrumental rationality* makes it impossible for us to make money pump arguments anyway. And so if there is any hope for money pump arguments in favour of acyclicity, we cannot rely on *preference-based instrumental rationality* in our response to the challenge resolute choice poses. We proposed a desire-based notion of instrumental rationality in the last chapter instead. What defenders of resolute choice need to show is that agents may have instrumental reasons to be resolute according to this desire-based notion of instrumental rationality.

The more general challenge that resolute choice poses to money pump arguments is that it may be possible to avoid being money pumped while generally keeping one's cyclical preferences. One could

[2] McClennen (1990) expresses the same idea by requiring that sophisticated agents abide by what he calls 'separability' (p.122). That is, sophisticated agents treat continuation trees within dynamic decision problems like new trees.

either do this by acting counter-preferentially at specific points in a decision problem, or by adjusting one's preferences temporarily within dynamic choice contexts only. The two different interpretations of resolution (or unified choice) we presented here adopt these two alternative strategies respectively. But being resolute may not be the only way of exploiting these strategies successfully in response to money pumps. What I want to investigate now is in how far desire-based instrumental rationality allows for such alternative responses to money pump arguments.

4.3 Preference as a Summary of Desire

According to desire-based instrumental rationality, desires over simple states of affairs are the fundamental conative attitudes against which we judge the instrumental rationality of our actions. This raises the question of what the preferences over outcomes that feature in formal decision theories are in the light of these desires, and in what relation they stand to desire. In particular, are agents still required to be guided by their preferences in action once we adopt desire-based instrumental rationality?

Most philosophers and decision theorists regard preferences to be conative attitudes. Indeed, *preference-based instrumental rationality*, which we have said is wide-spread, regards preferences over outcomes as primitive conative attitudes. But now that desires over simple states of affairs are also in the mix, can we still interpret the preferences that standard decision theory appeals to as conative attitudes? If yes, how do they relate to desire?

One first alternative is to say that preferences are simply alternative conative attitudes that stand in no special relation to desire. But this precludes from the start that we should be able to give an instrumental justification for having certain preferences or acting in accordance with them. If actions are judged by how well they serve our desires, and preferences stand in no special relation to desire, then we seem to have no reason to let our actions be guided by our preferences in any way. The conclusion to draw from the money pump scenarios would then simply be that the agent is wrong to let her actions be guided by her preferences in the first place. Moreover, we can't conclude from the instrumental failure of an agent who is money pumped that anything was wrong with her preferences.

The more plausible picture is one whereby preference stands in a closer relation to desire. Above, we said that we ordinarily think that our desires *explain* our preferences over outcomes: I prefer Apartment A over Apartment B because I desire a beautiful view and a short commute, and take this to outweigh the fact that A is smaller. And the self-torturer prefers going up by one setting because he desires to have more money. Perhaps, then, we can understand our preferences over outcomes as expressions of all of our desires that are relevant for the outcomes under consideration taken together, as a result of the difficult weighting process alluded to in the last chapter. Our preferences could be understood as comparative evaluations of outcomes on the basis of our desires over all the states of affairs the outcomes comprise, that is, as summaries of our desires.[3] Let us first consider such an account of preference.

Can we justify *maximization* and *acyclicity* for preferences understood in this way? In the case of *maximization*, we can already note one caveat here. Unless we take preferences to be infallible as

[3] Heath (2008) similarly regards both desires and preferences as models for our diffuse affective states. This for him grounds a kind of cognitivism about desire and preferences, whereby desires and preferences are similar to beliefs, by being answerable to how well they represent those diffuse states. We are not committing to cognitivism about desire about simple states of affairs. However, on the present notion of preference, we could understand preferences over outcomes as a kind model of the agent's diffuse desires. This makes preferences answerable to desire, and rationally criticizable if they are a bad model — that is, don't reflect the agent's desires well. The agent's desires themselves, however, are not similarly criticizable by a standard other than themselves.

summaries of the agent's desires, then it can happen that the agent's preferences misrepresent her desires. Indeed, if we want to be able to give the kinds of instrumental justifications for resisting temptation we suggested at the end of Chapter 2, we have to embrace such fallibility. Since the kinds of mistakes that we suspected may be going on in temptation cases seem very familiar, and since we experience the process of forming preferences in the face of conflicting desires as difficult, I take the assumption of fallibility to be very plausible.

Strictly speaking, we are thus entertaining a notion of preference whereby it is a binary evaluation of outcomes that ideally captures all the agent's relevant desires correctly. If our preferences over outcomes are fallible as summaries of our desires in this way, then *maximization* can hold at best as a conditional principle: Agents ought to maximize if their preferences represent their desires correctly. But as we already noted, this may be a caveat that decision theorists might be willing to take on board.

Can we justify *maximization* in its conditional form with appeal to desire-based instrumental rationality? Suppose an agent's preferences in fact correctly express all of her relevant diverse desires. It now seems like we can explain again why she may be rationally required to be guided by these preferences. As a matter of fact, the consequences of our choices are full outcomes: I choose to live in an apartment, along with all that that implies. I do not only choose a beautiful view. In fact many different desires seem to be relevant for my choice of apartment. A conative attitude over full outcomes thus seems more directly applicable to my choice than a desire for a simple state of affairs. And if my preferences capture all the different desires over the states of affairs the outcome comprises, then acting on the preference means that my action is still ultimately based on my desires. This seems to license the claim that desire-based instrumental rationality demands an agent's actions being in some sense guided by her preferences, if those preferences are indeed correct representations of her desires.

The next section argues that unfortunately, *maximization* cannot be given such a simple justification if we also want any hope of money pump arguments being successful. Ultimately, I will argue that we cannot justify the core principles of standard decision theory on the understanding of preference we are entertaining here. The most popular conception of preference amongst philosophers is thus of no use in justifying standard decision theory. The second part of this chapter will argue that an understanding of preference as something closer to action, as it is more common in economics, is more amenable to justifying the core principles of standard decision theory under certainty.

4.4 Uniqueness and Preference-Guidance

Maximization requires an agent to choose an outcome to which no other is strictly preferred. Schwartz's rule for intransitive preferences requires that an agent be similarly guided by her preferences in action. For instance, in a binary choice between two outcomes, where one of the outcomes is strictly preferred, both rules require the agent to choose the most preferred one. As we have seen, such a choice behaviour is required for the money pump arguments to get off the ground.

Does the characterization of preference we just gave provide a desire-based justification for a requirement to be guided by one's preferences in either of these two senses, at least when preferences correctly capture the agent's desires? Letting one's preferences guide one's actions seems to be rationally demanded under this conception of preference only if there is only one preference relation over outcomes that is admissible given the agent's desires.

Suppose that there is only one way in which the agent can synthesize her desires over features of outcomes into preferences over full outcomes, one preference relation that correctly summarizes her desires. For each pair of outcomes, the agent's desires over features of those outcomes completely determine whether the agent should prefer one over the other or be indifferent. Let me call this condition *uniqueness*.[4] If uniqueness holds, and the agent's preferences correctly capture her desires, it seems like, if the agent fails to be guided by her preferences in action, she is not doing well by her desires.

Now suppose uniqueness fails: Several different preference relations would equally well express the agent's desires over features of outcomes. The agent only adopts one of these correct preference relations.[5] Now it seems like desire-based instrumental rationality does not demand that the agent is guided by her preferences. If she chooses outcomes in accordance with a preference relation that she does not have, but that would have been admissible given her desires, then she does not seem to be instrumentally criticizable according to desire-based instrumental rationality.

I here want to argue that if we want to give an agent with cyclical preferences reason to have acyclical preferences instead, then we cannot assume uniqueness. And thus, if there is to be any hope of justifying acyclicity with a money pump argument, we cannot make the straightforward argument for *maximization*, or preference-guidance more generally that we mentioned in the last section.

In order for there to be any hope of giving an instrumental reason for an agent to have acyclical preferences, given the notion of preference we are working with here, there must be an acyclical preference relation that actually expresses the agent's desires correctly. I will grant this here.[6] But now the question of uniqueness arises: Is this the one unique preference relation over outcomes that expresses the agent's desires correctly? I want to argue that this can't be so if we want to deliver a money pump argument.

If we do assume that uniqueness holds with regard to an acyclical preference relation, then the original cyclical preferences must have been a mistaken expression of the agent's desires. This is already a worrying implication, since it does not do justice to the intuitive plausibility of those cyclical preferences. But perhaps the money pump argument could then be understood as an argument that preferences that would express the agent's desires well would need to be acyclical. The reasoning could be that if the agent's preferences capture her desires fully, then actions guided by those preferences should not frustrate those very desires.

However, this argument does not work, since it takes for granted that agents should always be guided by their preferences in action. But we cannot do so in this case. If there is one unique acyclical preference relation that expresses the agent's desires well, we may be able to justify *maximization* with respect to

[4] There is a parallel debate in formal epistemology about whether there is a uniquely rational credence an agent may have given her total body of evidence. See, for instance, White (2005) and Titelbaum and Kopec (2016).

[5] Another possibility would be that in these kinds of cases, an agent should adopt a family of preference relations, expressing an 'imprecise preference'. Imprecise credence, too, is often represented by families of probability functions. See Bradley (2015). However, in that case, we would already have given up on *maximization*, since we would have to formulate new choice rules for families of preference relations. How we could then give an instrumental justification for each preference relation in the family being transitive is unclear to me. Most plausibly, a choice rule for 'imprecise preference' would be fairly permissive, and allow the agent to act in a way that is licensed by at least one of the permissible preference relations (that is, by maximizing with regard to the preference relation, or by following Schwartz's rule with regard to that preference relation). Principle Q would justify us saying that the agent should choose such that she is not money pumped, insofar as this is compatible with the permissive choice rule. But clearly, to conform to principle Q and such a permissive choice rule, the agent's individual preference relations need not each be transitive.

[6] There is some reason to be sceptical of this claim, especially in problem cases like the Self-Torturer Problem. It is not clear that in that case any acyclical preference order could express the agent's desires correctly. Such a preference relation would have to include a preference not to go up any further setting at some point if it is to keep the agent from self-torturing. But such a preference seems like a poor representation of the agent's desires to have money and be pain-free, given that the difference in pain between adjacent settings is not distinguishable to the agent when facing a pairwise comparison. However, I will conclude below that the better notion of preference to try and salvage the money pump argument is one that is closer to action anyway. And so while granting the assumption that some acyclical preference relation may represent the agent's desires correctly here, I am not committed to this claim.

those preferences. But the argument we just made assumes that the agent would also be guided by her preferences in action if she had cyclical preferences instead. And we cannot take that for granted. In particular, as long as we grant that some acyclical preferences would represent the agent's desires well, desire-based instrumental rationality would permit the agent to act in accordance with those hypothetical preferences, thereby avoiding being money pumped, while keeping the cyclical preferences.

Perhaps in some cases, we have independent reason for thinking that cyclical preferences misrepresent an agent's true desires. But given the intuitive plausibility of cyclical preferences in the problem cases we are dealing with here, we do not generally have such independent reason for thinking that cyclical preferences misrepresent the agent's desires. And in any case, we would then not be in need of money pump arguments. But it is widely agreed that money pump arguments are the best instrumentalist arguments in favour of acyclicity.

Hence, in order to be able to deliver a money pump argument in favour of acyclicity, we have to assume non-uniqueness at least in the sense that the original cyclical, as well as at least one acyclical preference relation would represent the agent's desires correctly. But this is also enough by way of non-uniqueness to show that money pump arguments fail according to the notion of preference we are dealing with, as the next section will argue.

4.5 Back to the Money Pump Argument

The last section showed that, if we want to deliver a money pump argument, we have to admit that in problem cases for transitivity, uniqueness fails at least in the sense that the original cyclical preferences, as well as at least one acyclical preference relation would correctly represent the agent's desires. This makes the requirement that the agent's actions should be guided by her preferences questionable. That is, it is questionable whether the agent should act in accordance with *maximization*, or alternatively Schwartz's rule when she has cyclical preferences.

If there are two preference relations that are correct expressions of the agent's desires, then it seems like, whichever the agent adopts, she is permitted by desire-based instrumental rationality to still act in a way that would be licensed by the other preference relation. And so whichever the agent adopts, it seems like she is permitted to choose counter-preferentially. We thus cannot assume preference-guidance anymore. What does this mean for the money pump argument? We said above that the money pump argument only gets off the ground if the agent is guided by her preferences in action.

Suppose the agent adopts an acyclical preference relation, she sticks by it over time, and maximizes with regard to it. In that case, she cannot be money pumped. Being money pumped is bad in terms of desire-based instrumental rationality. And so while we cannot give an independent justification for *maximization* anymore, one might think that money pumps now give a *joint* instrumental justification for acyclicity and *maximization*. After all, given stable preferences over time, acyclicity and *maximization* together guarantee that the agent abides by principle Q (Section 3.4) in the money pump scenarios we have been considering.

I will grant for now that the agent in fact sticks with whatever preference relation she adopts throughout the series of choices she is offered. Still, we cannot get a joint justification for acyclicity and *maximization*. Assuming stable acyclical preferences, maximization is a good way (though not the only way) of avoiding being money pumped. And assuming guidance by stable preferences, having these acyclical preferences is the only way of avoiding being money pumped. But we have no independent instrumental justification for either preference-guidance or acyclicity. This means that we cannot justify them jointly using a money pump. An agent can equally avoid being money pumped by keeping her cyclical preferences, and by refraining from always being guided by them in action.

In fact, we can understand the appeal to resoluteness, in its first interpretation, in this way. An agent who keeps her cyclical preferences, but manages to be resolute and to act against her preferences at some point, can also avoid being money pumped. As long as the resolute course of action is one where each individual action is permitted given the agent's desires, she will not violate Q. This will be so if the counter-preferential choice would have been licensed by some other permissible preference relation — perhaps the acyclical one the agent would adopt were she forced to adopt one.

The lesson Tenenbaum and Raffman (2012) draw from the self-torturer problem is along similar lines. They claim that, in the context of problems like the self-torturer problem, having cyclical preferences is perfectly reasonable. At the same time, they hold that agents are also instrumentally required to at some point — before they end up with a clearly undesirable outcome — act against their preferences. Like us, the sense of instrumental rationality they have in mind is clearly not preference-based, but takes instrumental rationality to be about more basic, and often vague ends. We can explain how the behaviour they describe can be consistent with desire-based instrumental rationality under the assumptions we have made. Since we have granted that there is some permissible acyclical preference relation that would correctly express the agent's desires, the agent does not seem instrumentally irrational if she chooses a stopping point consistent with that hypothetical preference relation.

We can give the same response in the apartment case. Suppose I get offered Apartments A, B, C, and A - ϵ in succession. One way of avoiding to end up with A - ϵ would be to form stable acyclical preferences over the options that rank A higher than A - ϵ, and maximize regarding those preferences each time I get to make a choice. But an alternative way would be to keep the cyclical preferences we described, and frustrate one of my binary preferences at some point before I end up with A - ϵ.

Why avoid being money pumped in the first way rather than the second way? Money pumps do not help us make that choice, and can thus not issue in a requirement to have acyclical preferences. In fact, if anything, on the notion of preference we have been dealing with, sticking with the cyclical preferences seems most natural, since those are the preferences that our agent intuitively starts out with as the most appropriate representation of her desires.

One may respond that part of the point of having preferences is action-guidance, and that we should thus go with the first response to money pumps of adopting acyclical preferences and letting those guide our actions.[7] However, either this argument relies on reinterpreting preference, or it is ineffective. On the one hand, we could redefine preference in such a way that we cannot choose to act counter-preferentially. For instance, we could say that preferences are dispositions to choose. But this would be a departure from the notion of preferences we have been dealing with here. It is in fact the alternative interpretation of preference we will consider below, and is popular particularly in economics in the guise of 'revealed preference theory'. As we will see, I take this notion of preference to be more promising, but it comes with its own problems.

On the other hand, we may mean that agents should form preferences over outcomes with the purpose of guiding choice in mind. Since only acyclical preferences can always be choice-guiding without

[7] That the point of preference is action-guidance is a common claim. Nozick (1993), for instance, writes, "[t]he function of preference, the reason evolution instilled the capacity for them within us, is to eventuate in preferential choice." (p.142) Andreou (2007) distinguishes between different notions of preferences and argues that it is 'preferences for the purposes of choice' that need to be transitive. I don't mean to suggest that these authors would subscribe to the argument I am describing here, however.

leading to the agent being money pumped, this may give the agent reason to form acyclical preferences. The problem with this argument is that we lack an instrumental justification for why agents should form preferences, in the sense we understood them, with the purpose of guiding choice in mind. That requirement looks a lot like *maximization* itself, and we just saw that we cannot justify that requirement instrumentally.

Perhaps the claim is just that agents in fact do form preferences with the purpose of guiding choice in mind. However, such a descriptive claim cannot ground the normative claim that agents *should* respond to money pumps by forming acyclical preferences and maximizing with regard to them. It would only establish that agents usually do that. However, first, cyclical preferences can be action-guiding outside of money pump contexts without leading the agent to be money pumped. Instrumental rationality is thus compatible with cyclical preferences fulfilling the purpose of action-guidance for the most part. And secondly, to the extent that cyclical preferences shouldn't rationally guide preference, the prevalence of cyclical preferences may just be evidence that agents do not form preferences with the purpose of always being action-guiding.

Thus, if we stick to a notion of preference that takes preference to be a comparative evaluation of outcomes on the basis of the agent's desires, then I don't think there is any decisive reason to avoid being money pumped in one way rather than the other here. And so money pumps fail to give a joint instrumental justification for *maximization* and acyclicity on this understanding of preference.

My argument has exploited a specific kind of non-uniqueness: If we want to make a money pump argument for the acyclicity of preference, we need to accept that the preference relations that express the agent's desires well are non-unique at least in the sense that the original cyclical preferences and one acyclical preference relation represent the agent's desires well. But then we cannot provide an independent instrumental argument for preference-guidance anymore, and money pump arguments also fail to jointly justify maximization and acyclicity.

Perhaps, however, we can at least provide an instrumental justification for either of the requirements if we are allowed to simply assume the other. Indeed, the money pump argument does seem to give a good justification for acyclicity if we simply assume that the agent is guided by her preferences in action. This does not appear to show much, however. Preference-guidance does not have any intuitive plausibility as a non-instrumental requirement of rationality, given that we have interpreted preferences as a conative attitude that has no conceptual connection to action. And given that we have adopted a desire-based notion of instrumental rationality, instrumental rationality also gives us no general reason to accept preference-guidance.[8]

Still, perhaps we can give an instrumental justification of *maximization* if we can assume transitivity. After all, we only had to assume the non-uniqueness we just appealed to because we wanted to offer a money pump argument in favour of acyclicity. For those who think that transitivity holds for other reasons there would still be value in showing that there is an instrumental justification for *maximization*

[8] Instrumental rationality may give an agent such reason if she has some desire that is served by letting her actions be guided by her preferences. If the agent had a desire to act in accordance with her preferences, then perhaps instrumental rationality requires her to adopt acyclical preferences and act in accordance with them in response to money pump arguments. However, while instrumental rationality cannot proclaim it irrational, I take such a desire to be highly unusual. There seems to be nothing desirable, per se, about being guided by one's preferences in action, according to the conception of preference we have been appealing to. And thus, even though appeal to such a desire may give an instrumental defence of acyclicity that is conditional on a desire, this is a defence will not apply to many actual agents. In that sense, I think it is different from the conditional defence of the requirements of standard decision theory we will provide below. Moreover, we have here bracketed concerns about the potential instability of preference. In fact, even granting a desire to be guided by our preferences in action, the money pump argument will still fail unless we can assume stability of preference, as argued in the second part of this chapter.

given transitivity. Standard decision theory would at least be partly instrumentally justified.

The next section argues, however, that we cannot even offer such a restricted instrumental justification of *maximization*. And that is because non-uniqueness holds more generally, especially in the kinds of problem cases for transitivity we considered. This non-uniqueness means that we cannot even give an instrumental argument for *maximization* if we assume transitivity. I will conclude that given desire-based rationality, interpreting preference as a conative attitude that expresses all the agent's desires makes it impossible for us to justify the core requirements of standard decision theory under certainty.

4.6 Uniqueness and Maximization

As we have argued above, under non-uniqueness, preference-guidance no longer follows directly as a requirement of desire-based instrumental rationality. If an agent acts counter-preferentially, but in a way that would have been supported by a preference relation that is an equally good summary of the agent's desires, she is not rationally criticizable according to desire-based instrumental rationality — unless, that is, she ends up violating principle Q (Section 3.4). The specific non-uniqueness we considered above stemmed from our desire to make a money pump argument for acyclicity of preference. Does non-uniqueness disappear if we are permitted to simply assume transitivity?

Here I want to argue that in the problem cases for transitivity we have been considering, *if* there is an acyclical preference relation that represents the agent's desires correctly, then it would be implausible if there were only one. That there is some acyclical preference relation that would summarize the agent's desires correctly is something that we have granted — indeed without that assumption, the money pump argument would be hopeless from the start, given the notion of preference we are dealing with here.

To start with the self-torturer problem, requiring that there is one unique acyclical preference relation that represents the agent's desires correctly would be requiring that there is only one unique acyclical preference relation in which the desires over being pain-free and having money find correct expression. Such preferences would identify both a first and a last permissible setting that the agent can reach before stopping. Suppose that the agent has the following preferences, for instance:

$$S_1 \prec S_2 \prec ...S_{289} \prec S_{300} \succ S_{301} \succ S_{302} \succ ...S_{1000} \prec S_1$$

And suppose that the agent's preference relation as a whole is acyclical. These preferences identify setting S_{300} as both the first and last permissible stopping point. If uniqueness holds, this must be the only preference relation that correctly captures the agent's desires. And then if the agent were to stop one setting earlier, or one setting later, this will mean that she will not have done well given her desires. She will have acted irrationally according to desire-based instrumental rationality. At the same time, if she stops precisely at S_{300}, she will not be instrumentally criticizable.

Given the structure of this decision problem, it is unlikely that an agent will be able to identify a precise first and last setting of which this will be true. This is due to the fact that the levels of pain of adjacent settings are indistinguishable to the agent when compared pairwise. Whichever setting her preferences pick out as the last permissible one to stop, it will seem to her that either, it will not have frustrated her desires unnecessarily to go just one setting further, or, that she has already gone too far given her desires.

Suppose then, that the agent forms acyclical preferences that imply a last permissible stopping point that she is sure is still permissible according to her desires, say S_{300}. Then it must be the case that she thinks that preferences that permitted her to go one step further would equally have been an expression of her desires:

$$S_1 \prec S_2 \prec ...S_{300} \prec S_{301} \succ S_{302} \succ S_{303} \succ ...S_{1000} \prec S_1$$

And then it will be difficult for us to justify *maximization* with regard to the preferences the agent actually does hold: Desire-based instrumental rationality seems to permit her to go just one step further, despite her preference not to.

For further evidence of non-uniqueness, consider the following variation: Imagine one day you are pressed to settle on a transitive preference order. It picks out S_{300} as the last permissible stopping point. The next day, you are again asked the same. Nothing about your underlying desires for money and the absence of pain has changed. You are not trying to actively reproduce your answer from the day before. Perhaps you forgot what you said before. Now I take it to in fact be unlikely that you would give precisely the same answer. And doing so seems to be permissible given your desires. There can be a change in what preference order you permissibly settle on without there being a change in your underlying desires.

As Tenenbaum and Raffman (2012) argue, the underlying reason for this is that the agent's end of being relatively pain-free is *vague*. Each time the agent faces the choice of whether to go up by one setting, it seems like her desire for being pain-free is not frustrated. Still, at some point, it is frustrated. The agent can simply not identify a binary choice that frustrates it. Tenenbaum and Raffman moreover argue that many of our ends or desires are vague: If I desire to write a book, then it is not clear precisely when I will have completed my project, and how many hours of procrastinating are compatible with my finishing my book by a particular deadline.

Tenenbaum and Raffman argue that cyclical preferences are appropriate in these cases. My point here is that if we grant that some acyclical preference relations may be appropriate representations of the agent's desires, these are not likely to be unique. Otherwise we will not have done justice to the underlying vagueness of the desires. But this means that we cannot require *maximization*. And so even if we think that the self-torturer should have acyclical preferences, the agent still seems to be permitted to act counter-preferentially.

Similar claims apply in the case of the apartment choice. Suppose I wanted to form acyclical preferences over the apartments that accurately reflect my desires over the features of these apartments. We can take certain things for granted, for instance that I will not prefer A - ϵ over A.[9] But how should I rank A, B, and C? I will have to give up at least one of the binary preferences that seemed initially plausible to me in order to construct an acyclical ranking of the apartments. But which one(s)? If there is one unique acyclical preference ranking that expresses my desires, then there will be one right answer to this question.

My desires might make it easy for me to construct partial rankings of the outcomes. In our example, for instance, I could arrive at three partial rankings, in terms of size, view, and commute respectively, as shown in Table 4.1. We already noted, in Section 3.6, that we sometimes speak about preferences between simple states of affairs. Since this seems to be equivalent to just a comparative notion of desire,

[9] Pettit (1991) even defines desiring a property of an outcome, such as my having money, as a disposition not to have this kind of preference: "To desire a property is to be disposed to prefer a prospect that has it, assuming that there is only one, among a set of prospects that otherwise leave you indifferent." (p. 153) But defining desires for properties in terms of preferences over prospects (or outcomes) seems to give up on desire-based instrumental rationality. It is better to treat this as a requirement of instrumental rationality akin to principles P or Q: If I desire money, then I should not desire an outcome where I have less money to an outcome where I have more money that is otherwise the same.

	Size	View	Commute
Apartment A	3	2	1
Apartment B	2	1	3
Apartment C	1	3	2

Table 4.1: Partial Rankings of Outcomes

we admitted it as part of the standard of instrumental rationality according to desire-based instrumental rationality. We usually only speak in this way, however, when considering states of affairs along some dimension. For instance, in our example, I prefer a larger apartment to a smaller apartment. I would not ordinarily say that I prefer the colour yellow to an extensive breakfast, even though I have desires for both of these things. Comparative desires along some dimension may enable me to construct partial rankings of outcomes like the one in Table 4.1.

We can also think of such partial rankings as resulting from individual desires being satisfied to a smaller or greater extent by different outcomes. If I desire a beautiful view, for instance, that desire will be satisfied to a greater degree by apartments with more beautiful views. And if I desire a large apartment, that desire will be satisfied more to the extent that my apartment is large. The same applies for the length of my commute.

These partial rankings still leave it open to me how I should aggregate them into a complete ranking. Social choice paradoxes such as Arrow's Impossibility Theorem, first presented in Arrow (1950), attest to the difficulty of constructing a social ranking out of a number of individual rankings. Similarly, it will be difficult to construct an 'all things considered' ranking out of the kinds of partial rankings just described. In fact, the cyclical preferences we started out with are the result of a simple majority vote between my desires, an illustration of Condorcet's Voting Paradox. Having partial rankings based on having my different desires have a majority vote will thus not work as a way to construct a transitive preference order over outcomes.

Of course, agents typically have more than partial rankings to go by. In particular, desires typically come in degrees. As we said before, making a decision in the context of conflicting desires is the result of some kind of weighting process. If desires come in degrees, then these should matter in this weighting process. But both the empirical and the theoretical literature on decision-making in the context of multiple dimensions of evaluation attest to the difficulty of making trade-offs in these choice situations, even if the agent can take strength of desires into consideration. This makes it unlikely that this process needs to result in one preference ranking that uniquely does justice to all the agent's conflicting desires.

To start with the empirical evidence, choices in multiple criteria contexts are especially vulnerable to framing effects, whereby an agent's preferences reverse when different options are presented to her. For instance, introducing further options that make one of the original options seem like a compromise makes the agent more likely to prefer that compromise option over the other options originally present.[10] Susceptibility to such framing effects is often taken to be irrational. However, desire-based instrumental rationality need not condemn it, if non-uniqueness holds. This offers us the more charitable interpretation of widespread behaviour.

What we seem to learn from the empirical literature on framing is that agents often form preferences 'on the fly' for a decision problem at hand.[11] Framing then occurs when different frames elicit different

[10] See Simonson (1989) for this example, and Tversky (1972), Huber et al. (1982) and Heath and Chatterjee (1995) for further examples of framing effects due to multiple dimensions of evaluation.

[11] See the volume *The Construction of Preference*, Lichtenstein and Slovic (2006).

preference relations. Instead of interpreting this as necessarily irrational, we could simply take it to be evidence of non-uniqueness in the face of various desires pulling in different directions. It could be evidence that the agent finds that several different preference relations would equally represent her desires, and finds it hard to settle on one. This would also mean that desire-based instrumental rationality does not condemn susceptibility to framing as irrational per se. If we want to be charitable in interpreting a widespread behavioural phenomenon, this gives us reason to accept that non-uniqueness may hold in these choice contexts.[12]

There is also a large formal literature in management and organization science on how to make choices in multiple criteria contexts.[13] This literature typically assumes that the decision maker in fact has transitive preferences over the various outcomes, but wants to come up with mathematical models that can predict or assist her in her choice. But the models could also be used by a decision maker who is yet to determine a preference relation over outcomes from her desires over features of outcomes. Understood in this way, the literature offers her a vast number of different ways of doing so, many of which have been incorporated into decision making software.

What is important for us here is that there is no consensus on which is the best method to use in any given choice context. According to one approach, for instance, the decision maker is asked to formulate a number of goals, and a decision is made by minimizing weighted deviations from those goals.[14] According to another approach, the decision-maker is asked to maximize a weighted sum of the values of various sub-criteria. Yet another approach has the decision-maker maximize a weighted product.[15] Various different methods for eliciting weights and sub-values are used. There is no guarantee that these different approaches will all yield the same choice recommendation, let alone the same preference relation over all options. This again appears to be evidence that non-uniqueness is likely to hold in multi-criteria contexts.

What is in question here is whether, if any transitive preference ranking represents the agent's desires well, the agent's desires determine a unique such ranking. The fact that coming up with a ranking of outcomes based on a variety of different desires is so difficult seems to speak for a negative answer in cases involving many different dimensions of concern. Especially when an agent starts out thinking that cyclical preferences are plausible, any transitive ranking of options will in some sense seem unnatural to her. Uniqueness would require that there is one such unnatural ranking that is right for her, given her desires. But that seems implausible.

What I have argued for here is that if there is a transitive ranking of outcomes that expresses an agent's desires, then this ranking is likely not going to be unique in the kinds of problem cases we considered, involving vague desires or multiple dimensions of concern. But this means that we no longer have an immediate justification for *maximization.* We cannot require that the agent be guided by her preferences simply because these express her desires. If several preference rankings would express the agent's desires correctly, then desire-based instrumental rationality appears to permit her to sometimes act against preferences she holds — as long as she chooses in a way that is endorsed by some preference ranking that would be permissible. For instance, in the self-torturer case above, even if the agent prefers

[12] Note that it also doesn't follow directly from desire-based instrumental rationality coupled with uniqueness that susceptibility to framing is irrational. However, to make it come out as rational given those assumptions, we have to assume that the agent's underlying desires in fact change between framing contexts. But that also does not seem to be a charitable interpretation. We do not usually take desires to be so volatile, and shifts in desire to be so unmotivated.

[13] See Habenicht et al. (2002) for an introduction and overview.

[14] See Jones and Tamiz (2010) for a textbook on this 'goal programming' approach.

[15] See Triantaphyllou (2000) on both the 'weighted sum' and the 'weighted product' models.

S_{300} to S_{301}, desire-based instrumental rationality seems to allow her to choose S_{301}.

One may worry that such permissiveness could again lead to an agent being money pumped. In the self-torturer problem, couldn't the agent use the reasoning we just employed to move up one further setting every time she is given a choice? As long as for each pair of adjacent settings, there is a permissible preference relation that licenses going up by one setting, it seems like, by the reasoning we just employed, the agent is permitted to go up all the way to the last setting, and then be money pumped. Similar claims apply to the apartment choice case. Does this mean that we can perhaps give a money pump argument not for acyclicity, which we have assumed, but for *maximization* this time?

We cannot, since the agent need not abide by *maximization* for every choice she makes in order to avoid being money pumped. For instance, if her preferences pick out S_{300} as the last permissible stopping point, the agent may act counter-preferentially by going up by one more setting. If she then stops, she will not have been money pumped. Desire-based instrumental rationality seems to say nothing against failures of *maximization* as long as they do not result in violations of principle Q (Section 3.4). As long as (1) the agent's action is licensed by some preference relation that would have been permitted given her desires, and (2) the action does not lead to the agent being money pumped, desire-based instrumental rationality can't say anything against acting counter-preferentially.

What all of this shows is that even if we were allowed to simply presuppose transitivity, desire-based instrumental rationality does not even give us a justification for *maximization* on the interpretation of preference we have been working with. Seeing that it also does not give us an instrumental justification for acyclicity, this interpretation of preference seems to make it impossible to justify the main principles of standard decision theory in the context of certainty in terms of desire-based instrumental rationality.

The only way to give an instrumental justification for standard decision theory in terms of desire-based instrumental rationality is thus to reinterpret what we may mean by 'preference'. This is what the rest of this chapter attempts to do.

4.7 Preference as Disposition to Choose

Rather than as conative attitudes over outcomes, preferences are sometimes understood behaviourally. Many economists, in particular, take choice to be revealed preference, and, in turn, preference to be hypothetical choice. In their defence of 'mindless economics', Gul and Pesendorfer (2008) describe the mainstream position as follows:

> In the standard approach, the terms "utility maximization" and "choice" are synonymous. A utility function is always an ordinal index that describes how the individual ranks various outcomes and how he behaves (chooses) given his constraints (available options). The relevant data are revealed preference data, that is, consumption choices given the individual's constraints. (p.7)

In the philosophical literature, too, preferences are sometimes understood as dispositions to choose. In the context of certainty, where each available action leads to one outcome for certain, we can understand the agent as choosing between outcomes directly. A binary preference would then be a disposition to choose one outcome over another, when both are on offer. The notion of disposition used here is usually extremely thin, and can again simply be understood as hypothetical choice.[16] Others have proposed

[16] See, for instance, Maher (1993), Chapter 1.

hybrid views. Gauthier (1987), for instance, takes rational preference to require both a disposition to choose, and the corresponding conative attitude.

Understanding preference as hypothetical choice may help us, since under that definition, the question of whether one ought to be guided by one's preferences in action appears to become otiose. To have a preference for one thing over another would simply mean that one would choose one over the other if offered the choice. Perhaps, then, the money pump argument in favour of acyclicity may go through after all. The hope is that when justifying *maximization* becomes unnecessary, acyclicity can be justified instrumentally with a money pump argument after all. This hope also includes hybrid views that take a disposition to choose to be necessary for a preference to exist.

As I mentioned above, most philosophers interpret preferences as conative attitudes rather than as dispositions to choose. A standard critique of behaviourist notions of preference, as brought forward by Hausman (2000) or Köszegi and Rabin (2007) is that we often take preferences to explain action. In order to do so, they should not themselves be defined behaviourally. However, this criticism loses much of its force when we acknowledge that on the proposal we are considering here, we are not endorsing behaviourism about decision theory more widely, as, for instance, Gul and Pesendorfer (2008) seem to be advocating it.

We are here considering standard decision theory to be a desire-based theory of instrumental rationality. Even if we view preferences as dispositions to choose, these dispositions to choose should be answerable to the agent's desires, such that the agent's desires ideally explain them. Moreover, any requirements on how preferences relate to each other, such as transitivity, should be shown to be helping the agent do well by her desires. Given that we can, on this picture, appeal to desires to explain the agent's actions, it may not be so hard to accept that preferences themselves may not be explanatory. And so I think the most common reason for rejecting a behaviouristic interpretation of preference need not concern us here.

There are some problems with the behaviourist definition of preference even for our purposes. For one, it makes it hard for us to characterize indifference.[17] Intuitively, it seems like one can be indifferent between two options, while still having a disposition to choose one over the other. There may be a fact of the matter as to which of two options you will choose if you have to choose, without it being the case that you strictly prefer one. I will set this problem aside here, however.[18]

On the present understanding of preference, preferences are dispositions to choose one outcome over another if both are on offer. There is also a question of how to understand the condition "when both are on offer" here. If we don't want to rule out intransitive preferences from the start, then we can't mean that preferences are dispositions to choose one rather than the other no matter what further options are also on the table. For one, if there are further options on the table, the agent may choose one of those options instead. We may deal with this by adding further conditions. For instance, we could say that a preference of a over b is a disposition to choose a when both are on offer, and no other available option is preferred to a.

Still, on this notion of preference, we cannot make sense of cyclical preferences. As we have seen, when choosing between a set of options over which you have cyclical preferences, you must frustrate

[17] This is the reason why Savage (1954) rejected this notion of preference. Maher (1993), Chapter 1 provides a more nuanced argument.

[18] While it may be impossible to keep indifference as part of the preference relation, nothing precludes us from characterizing it with appeal to the agent's underlying desires. We could say, for instance, that an agent is indifferent between a and b if and only if the agent's desires permit her to prefer either — that is, they permit her to choose either if offered the choice between them.

at least one of your binary preferences. But if preferences are dispositions to choose one option over another if both are available, then it can't have been true that you had that preference after all.

McClennen (1990) formulates conditions on choice, such that the agent's choice behaviour can be described by a pair-wise preference relation that satisfies *weak ordering*, and with regard to which the agent maximizes (pp.22-25). The condition that needs to be met is that choice is *context-free.*[19] Context-freedom basically means that an agent's choices over a set of options are not affected by adding further options, unless those further options are themselves chosen.[20] It requires that the agent is not susceptible to the framing effects involving adding further options that we mentioned earlier. If we define preference in a way that already presupposes that such a condition is met, then we are already guaranteed transitivity.

But that is exactly what we are doing when we think of preferences as dispositions to choose that are in a similar sense unaffected by what further options are on the table. We have seen that cyclical preferences are impossible under such a notion of preference. And so this notion of preference is unhelpful. Being guaranteed transitivity through our definition of preference would be great if we also thought that we in fact all have preferences in the sense of preference we are appealing to. But this seems to be precisely what the controversy is about. Sceptics about transitivity would simply respond that we need not have dispositions to choose in the strong sense this notion of preference requires. And even if we think that rationality requires transitivity, it seems implausible to say that the agents we described as having cyclical preferences in fact fail to have preferences in the proper sense.

The better approach is to think of a preference of a over b simply as a disposition to choose a over b, when only those two options are available. In that case, the question of whether an agent should be guided by her preferences in a binary choice becomes otiose. As we have seen, such preference-guidance in binary choices is implied both by *maximization*, and by Schwartz's rule, which also applies to cyclical preferences.

Still, note that now *maximization*, or preference-guidance more generally, when choosing amongst a larger set of options still needs justification. It is still an open question why an agent should maximize with regard to her binary choice dispositions when many options are on the table. I will get back to this question in Section 4.9. But for now, I will set it aside to turn back to the money pump argument. In the money pump argument, the agent is only confronted with binary choices.

With a notion of preference as binary choice disposition in hand, it is no longer open to us to say that an agent can avoid being money pumped by keeping her cyclical preferences, and at some point acting against her preference. In the self-torturer problem, for instance, by consciously stopping early, the agent reveals that she has changed her binary choice disposition. She at that point no longer 'prefers' to go up by one setting, according to the notion of preference we are here considering. By the time the agent stops at some setting, her preferences are thus no longer the same cyclical preferences we assumed she started with.

[19] I will follow McClennen in speaking of 'choice context' when talking about what options are available to the agent. I will refer to 'choice situation' when referring to the wider context in which a choice is made.

[20] Formally, choice is context-free if the following two conditions hold for the agent's choice function C, which specifies the subset of the options O that the agent deems choice-worthy (McClennen thinks that the choice function need not determine a unique choice-worthy option for each set of options, admitting also that the choice function need not be interpreted behaviouristically):

Alpha: A choice function C defined on O satisfies *Alpha* just in case for all a in S, and all $S*$ such that $S*$ is a superset of S, if a is not in $C(S)$, then a is not in $C(S*)$.

Beta: A choice function C defined on O satisfies *Beta* just in case for all a and b in S, and all S* such that S* is a superset of S, if both a and b are in $C(S)$, then either a and b are both in $C(S*)$ or neither is in $C(S*)$. (p.23)

Does this mean that the money pump argument is finally successful? After all, the possibility of acting counter-preferentially is what before enabled agents to avoid being money pumped while keeping their cyclical preferences. The next section argues that the money pump argument is in fact only successful for agents who aim to have choice dispositions that are stable over time.

4.8 Money Pumps and the Stability of Preference

If normative decision theory works with a notion of preference that identifies it with disposition to choose, then it is only a theory of instrumental rationality insofar as these dispositions to choose are formed with a view to fulfilling the agent's desires. For instance, if I desire money, then I should be disposed to choose Apartment A over Apartment A - ϵ.

Above, we argued that if we understand preferences to be comparative conative attitudes over outcomes that express all the agent's different desires over features of outcomes, and if we want those preferences to be acyclical, then uniqueness is unlikely to hold in all choice situations. Especially in the problem cases we considered, if there is one acyclical preference relation that expresses the agent's desires well, there will be several.

A parallel claim will be true for preferences as choice dispositions, if those choice dispositions are supposed to be based on the agent's desires. Several different preference relations understood in this way will be appropriately supported by the agent's desires. Earlier, non-uniqueness caused trouble for the money pump argument, because without it, we cannot justify preference-guidance. Now that preference-guidance is not an issue, at least not in binary choices, is non-uniqueness no longer problematic? Unfortunately, it still is.

Under non-uniqueness, adopting acyclical preferences only keeps an agent from being money pumped if she also sticks to one of the permissible preference rankings over time.[21] If she does not, then she may still be money pumped, even though she has acyclical preferences at each point in time, and even though her preferences at each point in time are appropriately based on her desires. In the apartment case, suppose that at the time I choose between Apartment A and Apartment B, and then between Apartment B and Apartment C, I have the following acyclical preferences:

$$\text{Apartment A - } \epsilon \prec \text{Apartment A} \prec \text{Apartment B} \prec \text{Apartment C}$$

I choose Apartment B in the first choice, and Apartment C in the second choice. By the time I am then offered Apartment A $-\epsilon$, however, I have the following acyclical preferences, which are equally allowed by my desires over the features of the apartments:

$$\text{Apartment B} \prec \text{Apartment C} \prec \text{Apartment A - } \epsilon \prec \text{Apartment A}$$

I choose Apartment A - ϵ. Thus, I was money pumped. If we have no way of ruling out that such shifts in preference occur, then money pumps need not provide us with an argument in favour of acyclical preferences. In this example, I would have ended up making the same choices had I had stable, cyclical preferences. Acyclical preferences guarantee that we can't be money pumped only if we also have preferences that are stable over time.

Similar claims apply to the self-torturer problem. An agent may have acyclical preferences, each acyclical preference relation implying a last permissible stopping point, at every point in time. But if

[21] As Reiss (2013) puts it (p.39), dynamically inconsistent agents can be money pumped even if they have transitive preferences.

this preference relation is not stable over time, the agent may still go all the way up to the last setting, and then be money pumped. This would be the case if the last permissible stopping point is never between two settings the agent is considering. The last permissible stopping point is never where the agent is looking, as it were.[22]

Now we may again want to give an argument that money pumps jointly justify stability of preference, understood as choice disposition, and acyclicity. After all, agents who satisfy both will avoid being money pumped, without ever choosing in a way that is itself problematic in terms of desire-based instrumental rationality.

However, this argument again fails, because agents may also avoid being money pumped by having preferences that are changing but cyclical. All we need for an agent to not be money pumped is for her to stop the trading at some point before she has lost money. Under the notion of preference we are now considering, she must at some point prefer to stick with an outcome rather than trade it for the next one offered. And it is enough if she has that preference at the time when she is money pumped.

In the self-torturer problem, for instance, the agent could have the cyclical preferences we stipulated for most of the time in the series of choices she is offered. That is, most of the time, it could be true that were she offered a binary choice between two adjacent settings, she would be disposed to choose the higher one. It just needs to be true that, when she is offered the series of choices in the self-torturer problem, then at some point, she will be disposed to stick with a lower setting. But even at that point in time, her preference relation over all settings as a whole need not be acyclical.

Suppose, for instance, that at the point in time when she is offered the choice between S_{300} and S_{301}, she is disposed to choose S_{300}. She could at the same time still have the following cycle in her preferences:

$$S_1 \prec S_2 \prec ...S_{300} \prec S_{302} \prec ...S_{1000} \prec S_1$$

These preferences are only minimally different from the cyclical preferences that seemed most natural to the agent. If anything, they are thus more accurately based on the agent's desires than fully acyclical preferences would be. They just include the minimal change necessary to keep the agent from being money pumped.

Of course, given the preference cycle still contained in the agent's preferences, the agent may still be money pumped when she is offered a different series of trades — namely one that leaves out setting S_{301}. But actually being money pumped would require that the agent's preferences remain stably what they are over that series of choices. Again, the agent could avoid being money pumped by adopting a crucial preference to not trade further at some point in time in that new series of trades. If we give up the idea that the agent needs to have stable preferences over time, and across different choice situations, then agents can avoid being money pumped without having fully acyclical preferences.

The second interpretation of the resolute dynamic choice strategy discussed earlier can be understood as just this kind of response to money pumps. Resoluteness there required adjusting one's preferences

[22] Raffman (2012) makes a related claim about the binary relation "appears the same as". She argues that just because any two neighbouring items in a perceptual continuum appear the same need not mean that "appears the same as" is non-transitive. This could be so if the extension of the relation was unstable in such a way that, whenever we look at two neighbouring items, they appear the same. A difference in appearance is never where the agent is looking. She also presents experimental evidence that supports the claim that this is what is going on when people are confronted with colour continua. My claim here is that just because the agent has transitive preferences at each point in time need not mean that she ever in fact has a preference to stick with the lower of two settings. If the agent's preferences can shift, the last permissible stopping point may simply never be where the agent is looking. Seeing that what the agent is faced with in the self-torturer problem is in fact a continuum of pain, the instability of "appears the same as" may well explain why an agent could form preferences that are unstable in a similar way.

within a dynamic choice problem only, and only insofar as is necessary to end up with the outcome one would choose in a single choice (that is, not money pumped). The lesson is that agents can keep their cyclical 'given' preferences for the most part, if they just make the necessary adjustments to their preferences needed to make sure one isn't money pumped. But these adjustments are specific to each dynamic choice problem one faces.

In a sense, this challenge to money pump arguments is similar to the one we encountered above. There, we said that an agent can avoid being money pumped by acting against her binary preferences at some crucial points in time. Now that we have defined preferences as dispositions to choose, agents can avoid being money pumped by having a (temporary) *preference* to not trade anymore at those same crucial points in time. In either case, the agent need not have fully acyclical preferences at any point in time to avoid being money pumped. Similarly, the two interpretations of resoluteness described above seem to capture essentially the same kind of response to money pump arguments.

We are thus left again with two ways of avoiding being money pumped. One can adopt stable, acyclical preferences, or one can adopt cyclical preferences that are unstable in just the right ways. Can we say anything in favour of avoiding being money pumped in one way rather than the other?

What may count in favour of adopting stable acyclical preferences is that we are often not in a position to know what further choices we will be offered, and we sometimes forget what choices we were offered in the past. If that is so, we might fail to adopt preferences that are cyclical but unstable in just the right way to avoid being money pumped. Suppose, for instance, that I do not make my preferences acyclical, and instead plan to adapt my preferences temporarily when needed to avoid being money pumped. Further suppose that, in the apartment case, by the time the rental agency makes me the offer of trading back to Apartment A for a $25 dollar fee, I forgot that I could have had Apartment A for free earlier. I might then fail to choose in a way that keeps me from being money pumped, and go for Apartment A, paying the fee. This would not have happened had I adopted stable acyclical preferences.

In general, a strategy that requires me to adopt a specific disposition to choose on the right occasion, specifically to avoid being money pumped, requires an agent to keep track of potential money pumps to try and avoid them. But often, this will be difficult. Adopting stable, acyclical preferences circumvents this difficulty. With such preferences, agents cannot just stumble into being money pumped.

However, this response is not open to us if we are strict about the assumption that we are now only considering choice in the context of certainty. Such certainty presumably also involves certainty about what choices we will be offered in the future, and certainty about what we will forget. And then the alleged advantage of having stable acyclical preferences disappears.

Moreover, there is a related disadvantage to the strategy of adopting stable, acyclical preferences. And that is that it requires an agent to have a specific disposition to choose even in situations where there is no danger of her being money pumped. And that may be unnecessarily restrictive. This comes out especially in the self-torturer problem. Suppose that the agent avoids being money pumped in the series of choices we described above by adopting acyclical preferences that have her stop at S_{300}. If those are her stable preferences, then the agent is presumably also required to choose S_{300} over S_{301} at a later point in time, in situations where she is only offered that one choice. But that seems unnecessarily restrictive. Much speaks in favour of choosing S_{301} in that situation.

In order to offer a money pump argument in favour of acyclicity, we need to decisively side with the strategy of adopting stable, acyclical preferences. And we do not have such decisive reason. At best, there are competing considerations that count in favour of adopting one or the other strategy of avoiding being money pumped. Desire-based instrumental rationality thus gives us no reason to adopt acyclical preferences that is independent of the content of the agent's desires. The project of providing purely instrumental justifications for the core requirements of standard decision theory in the context of certainty thus seems to have reached a dead end.

4.9 The Desire for Stability

On either understanding of preference we have been considering, there are alternative ways of avoiding being money pumped than adopting acyclical preferences and maximizing with regard to them. Moreover, these alternative responses are themselves consistent with desire-based instrumental rationality. We have thus failed to give a justification of *maximization* and acyclicity in terms of desire-based instrumental rationality. At least, we have failed to do so with a money pump argument, and without appeal to any specific desire the agent may have.

Here I want to argue that desire-based instrumental rationality could give us grounds for an instrumental argument in favour of acyclicity if the agent has a specific kind of desire. Suppose we understand preference as disposition to choose, and the agent has a desire that favours stability of this disposition both over time and across different choice contexts. In that case, money pump arguments seem to give the agent reason to adopt acyclical preferences. And so, desire-based instrumental rationality could provide us with an instrumental justification for acyclicity that is conditional on the agent having a specific kind of desire.

What the money pump arguments highlight, given that we understand preference as choice disposition, is that an agent with stable preferences can only avoid being money pumped by having preferences that are acyclical. But we saw that agents can also avoid being money pumped by having cyclical preferences that are unstable in just the right way. A desire for stability of preference could clinch things in favour of responding to the possibility of being money pumped by adopting acyclical preferences.

We argued above that in problem cases for transitivity, if there is one acyclical preference relation that expresses the agent's desires correctly, there will be several. This causes problems, because it creates the possibility of the agent switching between different acyclical preference relations without directly misrepresenting her desires. If the agent has a desire for stability of preference, however, then once the agent has settled on a preference relation, this preference relation may then become the only relation that expresses the agent's desires correctly from then on.

For this to be so, the preference relation over outcomes needs to be responsive not only to the agent's desires over features of outcomes, but also to the agent's desires concerning her preference relation itself — such as the desire for stability. There is an interesting question here of how these may interact. But it does not seem mysterious that the agent could weigh up these different kinds of desires. Seeing that we are conceiving of preference as choice disposition here, the desire for stability of preference is not a second-order conative attitude, that is, a desire over an attitude that is itself conative. The desire for stability of preference is simply a desire to have stable choice dispositions. Such a desire can be treated as just another reason in favour of adopting one or the other course of action — in this case, the course of action that was favoured by the agent's previous choice dispositions.

So how could the desire for stable preferences act as a clincher? When different preference relations would be permissible insofar as the agent's desires over features of outcomes are concerned, this desire can decisively count in favour of the preference relation the agent held previously. If that is so, then the strategy of avoiding being money pumped by adopting cyclical preferences that are unstable in just the right way may perform worse according to the agent's desires than the strategy of adopting stable acyclical preferences. The agent will be frustrating her desire for stable preferences when another available course of action, which would have also been permissible given the agent's other desires, would have not done so. In fact, avoiding being money pumped without adopting stable preferences would thus still violate principle Q (Section 3.4).

Given the understanding of preference we are working with here, a desire for stable preferences is a desire for stable choice dispositions. An agent has a stable preference of a over b if she would choose a rather than b at any point in time, and in any situation when she is offered a choice between the two. Is there anything desirable about having stable choice dispositions in this way?

Many different desires are typically relevant for our choice between two fully described outcomes, and our choices are often made in the context of conflicting, or vague desires. These contexts tend to create the kind of non-uniqueness we argued for above. That is, there will be several permissible ways for the agent to trade off the relevant desires, resulting in several permissible preference relations. If an agent stably sticks to one preference relation, we can regard her as having settled on a way of trading off competing desires. The desire to have stable preferences, that is, stable choice dispositions, could be understood as a desire to be settled in this way. There seems to be at least something attractive about that.

The presence of a desire for stability of preference would have two further advantages. The first is that given such a desire, money pump arguments will be convincing even to those who are worried that given non-uniqueness, rationality cannot require that an agent not make losses over time. This worry is based on the fact that, given non-uniqueness, money pump scenarios can be structurally similar to the case of Abraham and his incommensurable choices described in Section 3.7.

If different preference relations are permissible given the agent's desires, then the agent's desires over outcomes still available to her at the time of action seem to permit her to switch between preference rankings. Any strategy we proposed for avoiding being money pumped restricts this freedom, at least at some points in time. They thus seem to require the agent to take into consideration more than her desires over outcomes at the time of action, which some believe goes beyond the requirements of instrumental rationality. If the agent has a desire for stable preferences, however, this worry does not apply anymore. The desire to have stable preferences can clinch things in favour of sticking with one's preferences. Once the agent has adopted one preference order, that preference order becomes the unique one that represents all the agent's desires.

The second advantage the presence of a desire for stable preferences would have is that a related desire could help us justify *maximization* even in cases where more than two options are available. On the notion of preference we adopted, an agent prefers a to b just in case she would choose a if she was offered only a and b. Given this notion of preference, the question of whether the agent ought to maximize in binary choices is no longer open. However, it is still an open question whether the agent ought to maximize in choices between more than two options.

This question is essentially about whether the agent's choice dispositions in binary contexts should carry over to a context in which an agent chooses between several options. We need to justify that the agent ought to abide by McClennen's *context-free choice*, that is, that her choices ought to be unaffected by the presence of options that are not themselves chosen. Does desire-based instrumental rationality have anything to say in favour of context-free choice? Unless the agent has a desire for her dispositions not to be affected by options that are not chosen, it does not seem so.

We may be able to defend *maximization* if uniqueness held. If the agent would choose a when both a and b are on offer, would choose a when both a and c are on offer, and would choose b when b and c are on offer, and these choices are the only ones that are appropriately supported by the agent's desires, it would seem odd if she didn't choose a when all of a, b and c are on offer, especially when these outcomes are described in a way that includes everything the agent cares about.

However, if different binary preferences would equally well have been supported by the agent's desires — say, $b \succ a$, $b \succ c$ and $a \succ c$ —- then it does not seem irrational for the agent to choose b when all of a, b and c are on offer. The agent could have equally had those preferences, and then *maximization* would have demanded that the agent choose b. There is no reason to think that desire-based instrumental rationality forbids that changing the alternatives on offer makes the agent choose in a way that is more consistent with binary choice dispositions the agent doesn't have, but could have had, given her desires.

Above, we argued that susceptibility to framing effects need not be irrational under desire-based instrumental rationality, because it may just be evidence of the agent switching between different permissible preference relations. The point here is that susceptibility to framing may not be irrational even if the agent sticks with one preference relation, and even if we understand preference as binary disposition to choose.

A desire to conform to context-free choice could change this. If the agent has a desire to be consistent in her choice dispositions in the sense that adding alternatives that are not themselves chosen doesn't change what alternative is chosen, then we could again defend *maximization.* It may be instrumentally irrational to be susceptible to the framing effect in question if one has a desire to not be susceptible to this framing effect. However, without such a desire, and given non-uniqueness, there seems to be nothing irrational about violating *maximization.* We need a desire for consistency of choice between different choice contexts.

Such a desire seems to be similar in nature to the desire to have stable preferences over time. In both cases, the agent wants her choice dispositions to be unaffected when something irrelevant in terms of her desires over outcomes changes — the passage of time, or a changing choice situation in the one case, and irrelevant alternatives offered, in the other. And so we can expect these two desires to come together. They are both instances of a desire to be settled in one's choice dispositions.

We have seen, thus, that money pump arguments in favour of the acyclicity of preference are only successful for agents who have a desire to have stable preferences, conceived of as choice dispositions, over time. Moreover, being able to appeal to such a desire comes with the two further advantages we just described. I also think that this kind of desire is not uncommon or strange.

Still, and importantly, instrumental rationality can require us neither to have the desire to have stable preferences over time, nor to have the desire to abide by context-free choice. This is because instrumental rationality was supposed to be silent about what desires an agent may have. At best, then, acyclicity can be justified with a money pump argument only conditionally. Money pump arguments show that instrumental rationality demands that agents who have a desire for stable preferences adopt acyclical preferences. For agents who lack such a desire, different responses to the money pump argument are equally compatible with instrumental rationality.

A similar claim seems justified for *maximization.* Reinterpreting preference as binary choice disposition only does away with the question of whether an agent ought to abide by *maximization* in binary choice contexts. A requirement for an agent to maximize in choice contexts with more than two options still needs justification. As we have seen, whether an agent ought to maximize in such contexts boils down to the question of whether she should follow context-free choice. But without a desire to abide by context-free choice, there seems to be no instrumental pressure to do so.

4.10 Conclusions

The two central requirements of standard decision theory in the context of certainty are *weak ordering* and *maximization*. If we want to understand standard decision theory as a theory of instrumental rationality, then we need to give instrumental justifications for both. I have offered what I believe to be the best case that can be made for these principles. Even this best case, however, only applies to agents who have a desire to have stable preferences over time, and preferences that are unaffected by choice context, where preferences are understood as choice dispositions. Standard decision theory, at best, only gives us conditional principles of instrumental rationality.

Maximization is usually taken to be an obvious requirement of instrumental rationality. And indeed, it seems to be well justified if preferences already form a weak ordering over outcomes, and we take preferences to be the fundamental conative attitude which forms the standard of instrumental rationality. But the last chapter argued that in order to give an instrumental justification for *weak ordering*, or at least acyclicity, we need to reject preferences over outcomes as the standard of instrumental rationality.

Instead, I argued, we need to adopt a desire-based notion of instrumental rationality. According to this notion, instrumental rationality requires an agent to do well by her desires over simple states of affairs. But on this notion, *maximization* becomes questionable. What this seems to show is that our past confidence in both *maximization* and *weak ordering* as requirements of instrumental rationality seems to be based on a fatal equivocation. On one way of understanding the standard of instrumental rationality, *maximization* is plausible, but our best arguments for *weak ordering* don't work. On another, the basis of those arguments for *weak ordering* is given, but *maximization* is no longer independently plausible.

Going with the desire-based notion of instrumental rationality, money pumps now seem to be charged with the double task of justifying both *maximization* and *weak ordering*. But we saw that they cannot live up to that challenge. Agents can avoid being money pumped by abiding by both, but they can equally well do so by violating both in the appropriate way.

We can avoid this problem somewhat by reinterpreting preferences as dispositions to choose, such that whether an agent should act on her preferences in binary choice contexts is no longer an open question. Here we saw, however, that we still face the problem that dispositions need not be stable. Agents may avoid being money pumped by having cyclical preferences that are unstable in just the right way. And so now the problem is that money pump arguments cannot jointly justify acyclicity and stability of preference.

The only agents for whom we can offer a clear instrumental justification for acyclicity, we argued, are agents who have a desire to have stable preferences, conceived of as choice dispositions, over time. Agents who have such a desire can be given an instrumental reason to adopt acyclical preferences. Moreover, positing a desire to have preferences that abide by context-free choice can help us justify *maximization* in non-binary contexts.

There is a certain irony in this conclusion. The most prominent alternative response to money pump arguments claims that agents can keep their cyclical preferences if they can behave in a way that is resolute. Resoluteness involves choosing a plan of action for a series of choices and then simply going through with it — either by acting counter-preferentially at the right times, or through isolated preference changes at the right times. Resoluteness is unpopular with many decision theorists. But if our argument is right, then justifying the core requirements of standard decision theory under certainty also only works if we can presuppose a kind of stability in agency. We need to presuppose that the agent desires to have stable choice dispositions over time. If anything, the stability that we need to assume agents strive towards is more thorough-going than resoluteness. Resoluteness only requires that an agent sticks by a plan of action for isolated dynamic choice problems. For money pump arguments to work, it needs to be the case that the agent desires to have stable preferences more generally.

Chapter 5

Decision under Uncertainty

5.1 Introduction

Most decisions we make are made in the context of uncertainty. We decide on careers, partners or on what city to live in not knowing if they will be right for us. Most everyday decisions also have uncertain consequences. On a smaller scale, we make decisions on whether to take an umbrella to work not knowing whether it is going to rain, on whether to buy concert tickets not knowing if we'll be able to find a date, and on whether to go to the concert not knowing if we'll like the music.

So far, we have ignored this uncertainty. However, its treatment of decision-making under uncertainty is arguably what standard decision theory is most famous for. The last chapters showed that the requirements of standard decision theory are already difficult to justify in the context of certainty. We argued that they can, in fact, only be justified for agents who have a desire to have stable choice dispositions over time and across different choice contexts. Here, we investigate whether the standard requirements of decision theory under uncertainty can be justified instrumentally. The most central such requirement, as we will see, is that of *separability*. This chapter shows that for parallel reasons, this requirement, too, can only be defended as a conditional requirement of instrumental rationality.

The standard instrumentalist argument that is made in favour of *separability* is a dynamic choice argument similar to the instrumentalist arguments we have encountered before. It aims to show that any agent who violates *separability* ends up making choices in some dynamic choice problems that leave her worse off by her own lights. This chapter argues that a common way of bringing out what is irrational about the agent who chooses in such a way understands instrumental rationality in the face of uncertainty in a way that means that the argument must fail. According to that notion, the agent's preferences over uncertain prospects form the standard of instrumental rationality.

Instead, I argue that the requirements of standard decision theory under uncertainty can only be justified instrumentally on a desire-based notion of instrumental rationality that takes the agent's desires to have only features of outcomes as their objects, not uncertain prospects or lotteries directly. This notion of instrumental rationality, however, seems to allow the agent to react to the dynamic choice argument in two ways: She could adopt separable preferences, or she could keep her non-separable preferences for the most part, and adjust them merely to avoid making a sure loss. Being resolute, as critics of expected utility theory defend it, is one way of doing so. As before, a desire to have stable preferences across different choice situations could help clinch things in favour of adopting separable preferences, and thus abiding by the central requirement of standard decision theory under uncertainty.

5.2 Expected Utility Theory

In decision-making under certainty, as we analyzed it so far, the agent is certain what the outcomes of the different actions open to her will be. Each action is thus associated with only one outcome. Under uncertainty, the agent does not know what the outcomes of her actions will be. There is thus no longer one outcome that is associated with each action. Each action may lead to a number of different outcomes.

Since Savage (1954), it has become standard to represent this kind of uncertainty by supposing that there are various different possible states of the world that are outside the control of the agent, but determine the outcomes of her action. We assume that there is a mutually exclusive and exhaustive set of states of the world $S_1 \ldots S_m$. Each action $A_1 \ldots A_n$ is then taken to lead to some assignment of outcomes $O_{11} \ldots O_{nm}$ to these states of the world. This can be represented in a state-outcome matrix like the one in Table 5.1.

	S_1	...	S_m
A_1	O_{11}	...	O_{1m}
...	...	...	...
A_n	O_{n1}	...	O_{nm}

Table 5.1: State-outcome matrix

An action, together with a state of the world, leads to some outcome for certain. But each action is associated with a distribution of outcomes over states of the world. These distributions are sometimes called *prospects*. Each row in the state-outcome matrix represents such a prospect. In the context of uncertainty, agents are not choosing between outcomes, but are usually taken to be choosing between prospects. I will thus speak of prospects and actions interchangeably. Decision theories under uncertainty typically require me to have preferences over these prospects as well as over outcomes.

To take a very simple example, suppose I am deciding how to get to work today. I could either cycle, or take the subway. Cycling is very pleasant if the weather stays nice, but very unpleasant if it rains. Taking the subway costs me $3, and is mildly unpleasant no matter what the weather is. A simple state-outcome matrix representing my decision problem could look as the one in Table 5.2.

	Rain	No rain
Cycle	Very unpleasant, but free	Very pleasant, and free
Take subway	Mildly unpleasant, and costs $3	Mildly unpleasant, and costs $3

Table 5.2: Should I cycle to work?

Here, the prospect of taking the subway is certain in the sense that what outcome I end up with, given that I cycle, does not depend on the way the world turns out. Taking the subway leads to the same outcome whether it rains or not. The prospect of cycling is uncertain in that what outcome I end up with if I cycle depends on whether it rains or not.

Decision theories in the context of uncertainty try to specify how agents ought to choose between prospects. While there are various different decision theories for the context of uncertainty, most of them are versions of *expected utility theory*. According to expected utility theory, agents ought to maximize their expected utility, or act as if they maximized the expectation resulting from some probability function (over states) and some utility function (over outcomes).

This is most easily explained if we think that we can take for granted that the agent can assign probabilities to the different states of the world, and can assign cardinal utilities to outcomes. Suppose, for instance, that I think that it is 50% likely that it will rain today during my commute. Moreover, I prefer the outcome of cycling when the weather is nice to taking the subway, and taking the subway to cycling when it rains. Suppose I can also determine that the first preference is twice as strong as the second. This may be expressed by assigning the following utilities to the outcomes:

$u(\text{pleasant cycling}) = 4$
$u(\text{subway}) = 2$
$u(\text{unpleasant cycling}) = 1$

Given these utilities and probabilities, we can calculate an expected utility for each of the prospects open to me by summing the products of the probability of each state of affairs and the utility of the outcome the prospect associates with it:

$EU(\text{cycling}) = 0.5 \cdot 4 + 0.5 \cdot 1 = 2.5$
$EU(\text{subway}) = 0.5 \cdot 2 + 0.5 \cdot 2 = 2$

In more general terms, the expected utility of act A_i is calculated as follows:

$$EU(A_i) = \sum_{j=1}^{m} p(S_j) \cdot u(O_{ij})$$

Expected utility theory requires agents to prefer acts for which this weighted sum is higher to acts for which this weighted sum is lower. It thus restricts what preferences agents may have over uncertain prospects, given their desires for outcomes, and their probabilistic beliefs. Expected utility theory adds to that the familiar requirement of *maximization*: Agents ought to choose an action such that no other action is strictly preferred to it. If the agent's preferences over actions track their expected utility, this comes to the requirement to choose an action for which expected utility is maximized. In our example, expected utility theory requires me to choose to cycle.

Note that if I had chosen different utilities that would have equally expressed the relative strengths of my preferences, the verdict would have been the same. This is why I am only required to have cardinal utilities, that is, utilities that are unique up to positive affine transformations to follow the advice to maximize expected utility. I can replace a utility function u with a utility function $v = a \cdot u + b$, where a and b are constants, and $a > 0$, without changing what action(s) expected utility theory recommends.

Still, a requirement to be able to assign cardinal utilities to outcomes is stronger than a requirement to be able to assign ordinal utilities, as was the case for choice under certainty. We saw in Chapter 2 that those who have a conception of utility as some real quantity that preferences ought to track may be able to justify the transitivity of preference easily. Appeal to such a real quantity also brings with it cardinality.

We have already noted that, in order for standard decision theory to be an instrumental theory of rationality, the quantity in question would have to express the degree to which an outcome is desired. The doubts we have raised about transitivity in the last chapters should make us sceptical that there is such a quantity of desiredness that gives us not only transitivity, but also cardinality. However, here

I want to bracket these worries. And that is because, even if we can presuppose a cardinal utility over outcomes along with probabilistic beliefs that we can use to calculate an expected utility, the requirement that an agent ought to maximize this expected utility is in need of defence. Our question here is whether this requirement can be defended instrumentally.

It seems fairly uncontroversial that *if* there is a real quantity of desiredness that applies to outcomes, then, in the context of certainty, we ought to choose an outcome that maximizes it. But this is not so obvious in the case of expected utility maximization. In our cycling case, it is not obvious that instrumental rationality requires me to cycle. Even though the difference in desiredness between cycling in good weather and taking the subway is larger than the difference in desiredness between the subway and cycling in bad weather, I may still prefer to play it safe. Taking the subway guarantees me that I end up with an acceptable outcome for sure. I may not want to take the risk of ending up with a lower utility.[1]

Defenders of expected utility theory who rely on a realist interpretation of utility need to argue that it is instrumentally irrational to be risk averse with regard to their realist notion of utility itself. For instance, they need to argue that, if I am contemplating whether to accept a 50/50 gamble, I am only allowed to play it safe and reject the gamble if my strength of preference of the status quo over losing is greater than my strength of preference of winning over the status quo. Since it is at least not obvious that this is so, something like the dynamic choice argument provided below is still necessary on a realist interpretation of utility.

So far, we have assumed that we can simply presuppose probabilities over states of the world, and cardinal utilities over outcomes. However, most decision theorists think that either just utility or both probability and utility are mere constructs that can be used to represent the agent's preferences over prospects as expected utility maximizing. Various representation theorems are supposed to show that such a representation is possible if the agent's preferences abide by a number of axioms. Under these interpretations, the normative requirements of standard decision theory are usually taken to be that one's preferences ought to abide by the axioms of one's favourite representation theorem. If one then maximizes with regard to them, one behaves as if one were maximizing an expected utility function. One chooses such that there are a utility and a probability function such that one's actions maximize the resulting expected utility.

While I will not go into the details of the various different representation theorems that have been proposed, what I here want to highlight is that there is one central requirement that is characteristic of all versions of expected utility theory. And that is the requirement of *separability*.[2] It finds expression in the axioms of various representation theorems, and is also implicit in the realist requirement to maximize expected utility.

Roughly, the idea behind separability is that an agent's preferences over two prospects should not be affected by what happens in states of the world that are not part of those prospects. If we think of preferences as assignments of value to outcomes and prospects, then we may say that separability requires there to be an independence in value of prospects over distinct states of the world.

Such a separability condition is in large part responsible for the possibility of an expected utility representation of an agent's preferences in the various representation theorems. And the utility repre-

[1] See also Buchak (2013), Chapter 1 on this argument.

[2] Another central requirement is that not only the agent's preferences over outcomes have to be transitive, but also her preferences over prospects. The last chapters were only concerned with justifying the first kind of transitivity. However, agents with cyclical preferences over prospects also stand to be money pumped. So our previous arguments apply *mutatis mutandis.*

sentation itself has an important separability feature as well. In expected utility theory, the overall value of an action is represented as a probability-weighted sum of the utilities of the outcomes occurring in separate states. This means that the value contribution of an outcome in one state will be independent of the value contribution of an outcome of another state, holding the probabilities fixed. The same holds for sub-prospects, consisting of assignments of outcomes to subsets of the states of the world featuring in the full prospect.

In the representation theorem introduced by von Neumann and Morgenstern (1944), separability finds expression in the independence axiom. This representation theorem takes probabilities for granted, but derives cardinal utilities.[3] If we already have probabilities over states of the world, then we can simply associate each outcome with the probability of the state of the world in which it occurs. The resulting probability distribution over outcomes is usually called a (simple) *lottery.* Von Neumann and Morgenstern's decision theory takes lotteries, which are either such simple lotteries, or probability distributions over other lotteries, to be the object of preference. When we assume that we can assign precise probabilities to each of the possible outcomes of our actions, then we can think of our actions as such lotteries.

Let $\mathscr{L}$ be the space of lotteries over all possible outcomes. The independence axiom then requires the following:

Independence: For all $L_x, L_y, L_z \in \mathscr{L}$ and all $p \in (0,1)$, $L_x \succcurlyeq L_y$ if and only if $p \cdot L_x + (1-p) \cdot L_z \succcurlyeq p \cdot L_y + (1-p) \cdot L_z$.

Independence claims that my preference between two lotteries will not be changed when those lotteries become sub-lotteries in a lottery which mixes each with some probability of a third lottery. For instance, suppose I know I get to play a game tonight. I prefer to play a game that gives me a 10% chance of winning a pitcher of beer to a game that gives me a 20% chance of winning a pint of beer. The independence axiom says that this preference will not be affected when the chances of me getting to play at all today change. The possibility of not playing at all tonight should not affect how I evaluate or choose between my options in the case that I do get to play.

In Savage's (1954) representation theorem, separability is expressed by the sure-thing principle. Savage's representation theorem does not take probabilities for granted, but also derives them as part of the representation of the agent's preferences over prospects. To state the sure-thing principle, we need to define a set of events, which are disjunctions of states. Let $A_i(E)$ be the act A_i when event E occurs. The sure-thing principle then requires the following:

Sure-thing principle: For any two actions A_i and A_j, and any mutually exclusive and exhaustive events E and F, if $A_i(E) \succcurlyeq A_j(E)$ and $A_i(F) \succcurlyeq A_j(F)$, then $A_i \succcurlyeq A_j$

The idea behind the sure-thing principle is that an agent can determine her overall preferences between acts through event-wise comparisons. She can partition the set of states into events, and compare the prospects of each of her acts for each event separately. If an act is preferred given each of the events,

[3] Von Neumann and Morgenstern's representation theorem is often understood to operate with an objective notion of probability, where probabilities exist objectively in the world, and do not represent the agent's degree of belief. In contrast, the representation theorems provided in Savage (1954) or Jeffrey (1965/1983), are usually understood as featuring subjective probability, which is supposed to measure the agent's degrees of belief. However, von Neumann and Morgenstern's representation theorem is in fact compatible with any interpretation of probability. All we need is to already have access to the relevant (precise) probabilities when applying the representation theorems. If we think of probability as the agent's subjective degrees of belief, we already need to know what those subjective degrees of belief are. If we think of it as objective chance, we need to already know what those objective chances are.

it will be preferred overall. That is, if a particular act is preferred no matter which event occurs, then it is also preferred when the agent does not know which event occurs. We can think of this as saying that there cannot be complementarities in value between sub-prospects over distinct states of the world. This independence in value is what makes this another expression of separability.

Assuming separability for preferences in the way that the independence axiom and the sure-thing principle do is essential for ensuring that there is an expected utility representation of the agent's preferences. Separability, in each of its guises, thus ensures that the utility representation has the important separability feature we already mentioned. If separability is problematic, it is thus problematic independently of any representation theorem, as well as for realists. In the following, I will consider separability mostly in the guise of the sure-thing principle. However, most of what I will say will apply to separability in all its guises.[4]

The next section introduces a famous apparent counter-example to expected utility theory that puts separability into question. We then consider what I take to be the most powerful defence of separability as a requirement of instrumental rationality.

5.3 The Allais Paradox

There are a number of famous examples that motivate the view that violations of separability are not in fact irrational, as expected utility theory claims they are. One such example is the Allais Paradox, as first presented in Allais (1953). The Allais Paradox runs as follows. First a subject is offered a choice between $1 million for certain on the one hand, and an 89% chance of winning $1 million, a 10% chance of winning $5 million, and a 1% chance of winning nothing on the other. Many people choose $1 million for certain when offered this choice. Next, the subject is offered the choice of either a 10% chance of $5 million, and nothing otherwise on the one hand, or an 11% chance of $1 million, and nothing otherwise on the other. Here, most people pick the first lottery, that is, the lottery with the higher potential winnings.

While in this formulation the choice problem has the agent choose between lotteries, we can easily translate it into a problem about prospects, if we assume that what the agent gets is decided by a random draw from 100 lottery tickets. The combination of preferences just described, now understood as preferences over prospects, seems sensible. We will henceforth call them 'Allais preferences'. However, they in fact violate the sure-thing principle, given a natural specification of the outcomes involved. This becomes evident when we represent the two choices in decision matrices, as in Tables 5.3 and 5.4.

Choosing lottery B in the first choice, and lottery C in the second choice violates the sure-thing principle. To see that, note that in both choices, the two prospects to be chosen from are identical with regard to what happens if tickets 1 - 89 are drawn. And thus, according to the sure-thing principle, the only thing that matters for the overall assessment should be what happens if tickets 90 - 100 are drawn.

[4] Separability also features in other famous representation theorems. In the decision theory developed by Jeffrey (1965/1983), for instance, separability is expressed by the averaging axiom. In Jeffrey's decision theory, acts, states and outcomes are all propositions, and all objects of preference. The averaging axiom claims the following:

Averaging: If A and B are mutually incompatible propositions, and $A \succcurlyeq B$, then $A \succcurlyeq (A \text{ or } B) \succcurlyeq B$.

The averaging axiom claims that an agent cannot strictly prefer or dis-prefer a disjunction to any of the disjuncts. When the propositions involved are outcomes or prospects over distinct states of the world, this requirement, too, expresses the idea that there is an independence in value between what happens in separate states of the world. Knowing only that I will end up with one of two outcomes or prospects cannot be worse than ending up with any of the individual outcomes or prospects.

	Tickets 1 - 89	Tickets 90 - 99	Ticket 100
Lottery A	$1 million	$5 million	$0
Lottery B	$1 million	$1 million	$1 million

Table 5.3: Allais Paradox: First Choice

	Tickets 1 - 89	Tickets 90 - 99	Ticket 100
Lottery C	$0	$5 million	$0
Lottery D	$0	$1 million	$1 million

Table 5.4: Allais Paradox: Second Choice

But for these tickets, the first choice, between lottery A and lottery B, and the second choice, between lottery C and lottery D, are identical. And so, the agent should choose lottery B in the first choice if and only if she chooses lottery D in the second choice.[5] Nevertheless, choosing lottery B in the first choice and lottery C in the second choice is both common[6] and does not seem intuitively irrational.

One way of reconciling these preferences with expected utility theory may be to argue that the outcomes are under-described by merely the money amounts that the agent will win following some draw of the lottery. Perhaps the outcome of receiving $1 million is different in the different prospects in a way that the agent cares about. There are various possibilities here. Perhaps, for instance, the agent cares about certainty, or she cares about avoiding regret or disappointment. And then, whether the $1 million was achieved as part of a certain prospect, and whether the agent will experience regret or disappointment should be part of the description of the outcomes in these decision problems.[7]

However, re-describing the outcomes to take account of disappointment, regret, or security arguably cannot do away with the violation of separability in the Allais Paradox. Weber (1998) provides an extensive argument to that effect. In any case, even if aversions to or desires for these things could explain why most people have Allais preferences, I think we can still conceive of an agent who desires nothing but money, and who still has the Allais preferences. Expected utility theory would declare such an agent irrational. But it is at least not immediately obvious that such an agent would be instrumentally irrational.

What can we say in favour of separability to such an agent, then? It has been pointed out that agents with Allais preferences, or indeed any agents who violate separability, are prone to making choices in dynamic decision problems that leave them somehow worse off in their own lights, or otherwise prone to behaving in a way that is instrumentally criticizable. The next section turns to such a dynamic decision problem.

[5] Similar reasoning applies for independence, if the agent is choosing between the lotteries that result if she assigns a 1% probability to each ticket being drawn and these probabilities are independent. We can then regard each lottery she is choosing between as a compound lottery of the sub-lotteries involving tickets 1 - 89 and 90 - 100 respectively. We then see that the Allais preferences must be violating independence. Moreover, no assignment of utilities and probabilities to the outcomes and states involved here exists such that the Allais preferences come out as expected utility maximizing. Thus, if the agent does assign cardinal utilities to outcomes, then she can't be an expected utility maximizer.

[6] See, for instance Morrison (1967) for experimental evidence that many people choose in this way.

[7] Graham Loomes and Robert Sugden have explored both making regret and disappointment part of the description of outcomes — regret in Loomes and Sugden (1982), and disappointment in Loomes and Sugden (1986).

5.4 The Dynamic Allais Problem

Hammond (1988) and Seidenfeld (1988) argue that for any preference relation that violates separability, there are dynamic choice settings where the agent will make counterintuitive or unacceptable choices. In these choice settings, choices are made consecutively as uncertainty is gradually resolved. The problems these authors point to arise because in such settings, sub-lotteries or sub-prospects that independence and the sure-thing principle require to be separable can be de facto separated in the dynamic structure of the decision problems, as agents decide about different sub-lotteries and sub-prospects gradually over time. And this can lead to patterns of choice that the agent can allegedly be instrumentally criticized for.

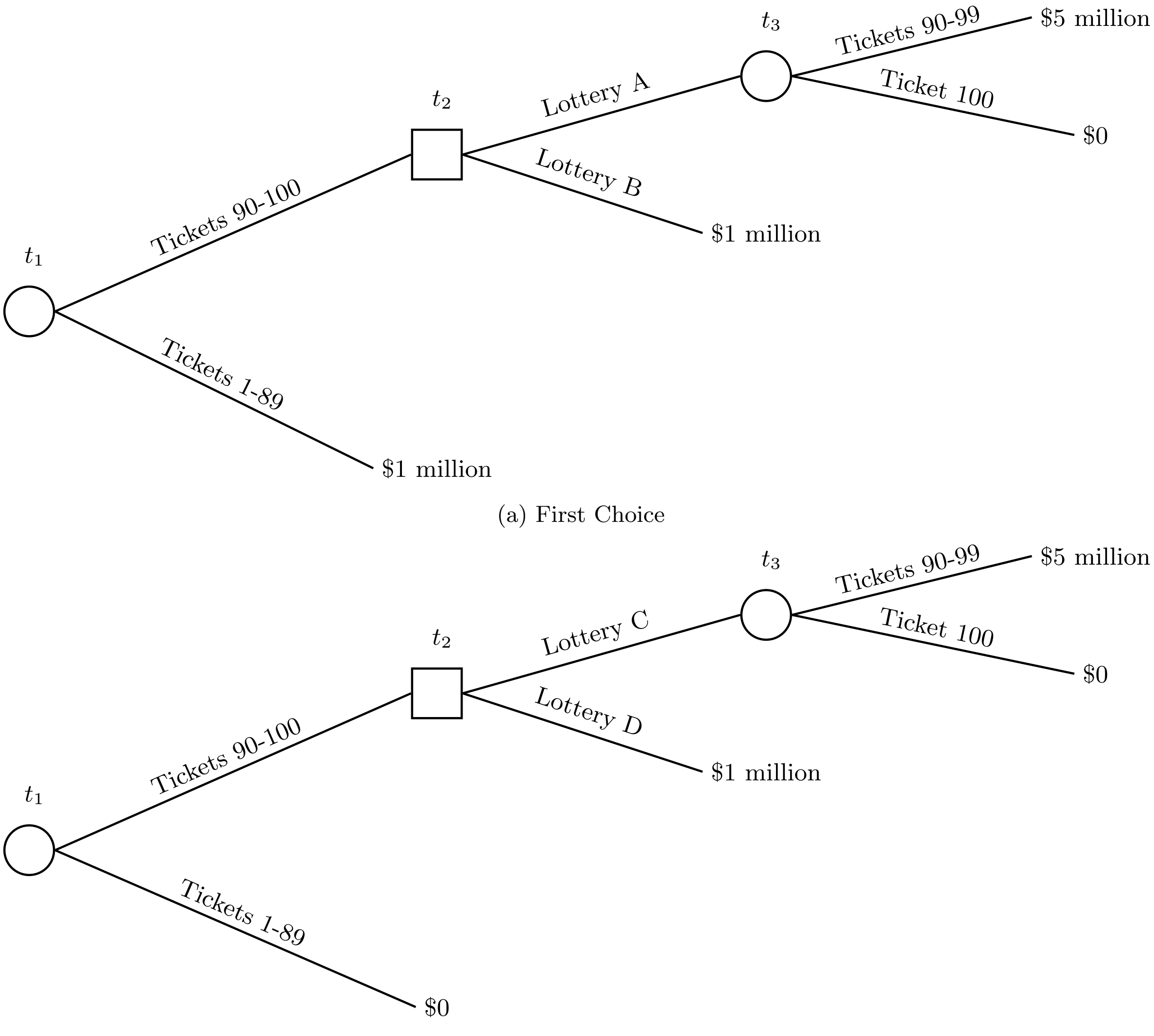
We can illustrate this with the following dynamic version of the Allais paradox, adapted from Machina (1989). In this dynamic version, agents only get to make a decision about which of the Allais prospects to choose after some of the uncertainty has already been resolved. They make a choice after they have found out whether one of tickets 1-89 has been drawn, or one of tickets 90-100 has been drawn. This can be illustrated in a decision tree, as in Figure 5.1. As before, the square nodes are choice nodes, at which the agent gets to decide whether to go 'down' or 'up'. The round nodes are chance nodes, at which chance decides between the branches. Let t_1 be the time at which the uncertainty is resolved about whether one of tickets 1-89 has been drawn, or one of tickets 90-100 has been drawn. t_2, in turn, is the time at which the agent gets to decide. At t_3, the agent finds out which ticket is drawn.

The interesting feature of this dynamic decision problem is that at the time when the agent gets to make a decision, at t_2, the rest of the tree, sometimes called the 'continuation tree', looks the same for the first and second choice, just as the sub-lotteries involving tickets 1- 89 we looked at in the static Allais problem. We might think that this means the agent should decide the same in both cases. If the agent strictly prefers the prospect of receiving \$1 million for sure to the prospect that gives her a 1/11 chance of winning \$5 million, she should prefer to go 'down' in both cases. If she has the opposite strict preference, she should go 'up' in both cases. If she is indifferent between those prospects, we can add a sweetener to one of the options to create a strict preference. This would presumably not alter the fact that she has the Allais preferences over the original full gambles with that sweetener added.

Suppose that in the continuation trees, an agent who has a strict preference between the prospects these trees lead to, call him Rory, prefers to go 'down' and get \$1 million for certain. He chooses in accordance with that preference. In that case, he will have chosen, over the course of the dynamic choice problem, to undergo lotteries B and D from the original problem. In the first choice, he receives \$1 million for certain in the course of the dynamic choice problem. In the second choice, he will run an 89% chance of receiving nothing, and an 11% chance of receiving \$1 million. But note that, in the second case, this is not a lottery Rory would have chosen at the beginning of the decision problem, were he to make a choice upfront.

More generally, we can say that if an agent must treat like continuation trees alike, then she will end up choosing in accordance either with lotteries A and C respectively, or with lotteries B and D respectively, but not according to the Allais preferences. That in turn means that for at least one of the choices, an agent with Allais preferences will end up choosing contrary to what she would have preferred at the beginning of the decision problem, before any uncertainty has been resolved.

Similar dynamic choice problems can be constructed for any preference relation over prospects that violates separability. In all of these decision problems, sub-prospects over which the agent has non-separable preferences are de facto separated in the dynamic structure of the decision problem, by resolving



(a) First Choice

(b) Second Choice

Figure 5.1: Dynamic Allais Problem

some of the uncertainty involved in the original problem before the agent gets to make a choice.

In such dynamic choice problems, agents like Rory end up acting, in the course of the dynamic decision problem, against their preferences over the prospects available initially. This has been held to be rationally problematic. According to an influential argument, both the requirement to act in accordance with one's preferences over the sub-prospects one faces, as well as the requirement to behave in dynamic decision problems like one would were one to settle on a course of action in advance are requirements of instrumental rationality. Rory violates the latter. And if he were to conform with the latter, he would violate the former. In fact, it can be shown that if we add a number of more technical assumptions, agents who abide by both of these requirements must be expected utility maximizers. However, the next section argues that there is no understanding of the standard of instrumental rationality according to which agents are required to abide by both of these requirements.

5.5 Consequentialism

As we have seen, for Rory the dynamic structure of the decision problem clearly makes a difference to what he will choose. If Rory were able to make a choice and stick to it before any of the uncertainty is resolved, and he would choose in accordance with his preferences over the prospects available then, he would choose in accordance with his Allais preferences. But in the dynamic decision problems, the prospect he ends up with in the second choice problem is not endorsed by his Allais preferences. However, it is claimed, for instrumentally rational agents, the dynamic structure of a decision problem should not make a difference in this way.

The dynamic structure of a decision problem not making a difference is part of what Hammond (1988) considers to be *consequentialist* decision-making — decision-making with an eye to the consequences of one's actions only. The thought is that if an agent's choice is changed by the dynamic structure of a decision problem in cases where the attainable consequences are the same, the agent's choice must have been influenced by something other than the consequences of her actions. In his reconstruction of Hammond's argument, McClennen (1990) calls this the requirement of normal-form/extensive-form coincidence: In dynamic choice problems, the agent should choose the same as she would, were she to simply choose one course of action at the beginning of the decision problem.

McClennen and Hammond show that this requirement, together with the assumption that the agent is sophisticated with regard to his given preferences and given some technical assumptions, implies that the agent must be an expected utility maximizer, and thus can't have Allais preferences. Remember that sophisticated agents solve dynamic choice problems by a process of backward induction. They make a prediction of their choice at the last choice node, assuming that they will pick one of their most preferred options then. They then similarly make a prediction of their own behaviour at the second to last choice node, given their prediction of their behaviour at the last node, and so on. Both Hammond and McClennen assume that the preferences sophisticated agents act on are their given preferences — the preferences they would have were they to face the prospect in isolation — and not preferences that were adjusted for the purposes of the dynamic choice problem. In our example, Rory chooses in such a sophisticated way. And we have seen that he ends up violating consequentialism.

Hammond's proof shows that if both sophistication and consequentialism are requirements of instrumental rationality, agents like Rory should adopt separable preferences. But are both sophistication and consequentialism requirements of instrumental rationality? I will argue that each of these principles is attractive only under conflicting understandings of what the ultimate standard of instrumental rationality is. Insofar as we find both of them attractive, we thus equivocate between these two notions of the standard of instrumental rationality. I conclude that we can at most justify one of these principles as a requirement of instrumental rationality.

We said in Chapter 3 that sophistication in the context of certainty is just the continued application of *maximization* with regard to the agent's preferences over the outcomes still available to her, combined with her accurate prediction that she will maximize in the future. In this case, in the presence of uncertainty, sophistication is just the continued application of *maximization* with regard to the agent's preferences over the *prospects* still available to her. Sophistication requires her to, at each point in time, choose the prospect that she prefers out of the prospects available to her then. Consequentialism, on the other hand, requires the agent to engage in a series of choices such that the prospect she faces at the outset of the decision problem is the one, or one of the ones she prefers then.

Just like *maximization* in the certain case, both consequentialism and sophistication under uncertainty require the agent to choose in accordance with some preference she has. In both cases, the preferences in question are preferences over prospects. We said that *maximization* in the certain case can be justified straightforwardly if we take the agent's preferences over outcomes to be the standard of instrumental rationality. Similarly, one might think that consequentialism and sophistication can be justified if we take the agent's preferences over prospects to be the standard of instrumental rationality. Note that, just like in the case of *preference-based instrumental rationality*, this again requires us to view preferences as conative attitudes, rather than as something closer to choice.

Sophistication would be a straightforward requirement of rationality if, at each point in time, instrumental rationality required the agent to do well by her preferences over the prospects open to her at that point in time. Call this account of the standard of instrumental rationality *open prospects*. The problem with *open prospects* is that it apparently cannot explain why anything is wrong with Rory for failing to abide by consequentialism. At the time when he gets to make a choice, he chooses the prospect he most prefers, out of the ones available to him then. If those preferences are the only standard of instrumental rationality, then that's just what he should do.

To say that Rory is rationally criticizable for failing to abide by consequentialism would be to say that Rory is irrational for failing to make it the case that in the beginning of the decision problem, he faced the prospect he most prefers then. It is true that according to *open prospects*, in the beginning of the decision problem, Rory's preferences over all the prospects open to him then would form the standard of instrumental rationality for any choices he makes then. But at t_2, the standard is a different one. Rory's choices are then judged against his preferences over the prospects available to him then, and no longer against whether they make it the case that he previously faced his most preferred prospect.

In a way, given Rory's preferences over prospects, *open prospects* means that the standard of instrumental rationality is shifting. At t_1, it endorses taking the gamble and going 'up' at t_2, and at t_2, it endorses playing it safe and going 'down' at t_2. This is because at t_1, Rory prefers a prospect that results from going 'up' at t_2, and at t_2, Rory prefers a prospect that involves her going 'down' at t_2. At t_2, Rory dis-prefers the prospect that is the sub-prospect of the prospect she would have preferred at t_1.

This shift in what choice the standard of instrumental rationality endorses does not seem to be relevantly different from other cases of shifts in preference, like the temptation cases we considered in Chapter 2. We there argued that as long as the tempted agent's shifted preferences are the standard against which her actions are judged instrumentally rational, then there is in fact nothing instrumental rationality can say in favour of resisting temptation. This is because resisting temptation would require the agent to act in accordance with preferences she does not hold at the time of action. The same seems to hold here. If the agent's preferences over the prospects available to her at the time of action are the standard of instrumental rationality, nothing seems to speak in favour of making it the case that her earlier preferences over the prospects open to her then are satisfied.

One might think that the present case is different from temptation cases, since Rory retains his preferences over the prospects available to him at t_1 throughout. Even as he chooses the safe option at t_2, he at the same time keeps his Allais preferences over the original Allais gambles. While this is true, and is part of what makes this case so puzzling, this observation does not help us in justifying consequentialism as a requirement of instrumental rationality, given *open prospects.* According to *open prospects*, it is simply not instrumentally relevant that Rory retains his Allais preferences over the prospects available at t_1. What matters at t_2 is that he does well by his preferences over the prospects open to him then. And the full Allais prospects are no longer open to him at t_2.

Another rejoinder may be that, since consequentialism is a principle about entire courses of action, the rationality of each individual action was really meant to be judged by whether it forms part of the best course of action, as in the two-tier accounts of temptation considered in Chapter 2. However, as we also saw in Chapter 2, moving to a two-tier account does not help us if the standard of instrumental rationality is shifting. At t_1, the course of action that has the agent go 'up' at t_1 is best according to *open prospects*. But at t_2, the course of action that has the agent go 'down' at t_2 is judged best according to *open prospects*. Evaluating the rationality of the individual action according to whether it forms part of the best course of action for the decision problem as a whole thus does not help in justifying consequentialism, given *open prospects*.

Granted all of this, the requirement of consequentialism appears to be either redundant or at odds with instrumental rationality. If the agent's later preferences over sub-prospects agree with her earlier preferences over the entire prospects open to her in the whole dynamic choice problem — as would be the case for agents with separable preferences — then the agent already abides by consequentialism simply by being sophisticated. But if she has preferences like Rory's instead, then consequentialism would require her to choose against her preferences over the prospects open to her at t_2. And that would be instrumentally irrational, according to *open prospects*.

Hence, according to the conception of the standard of instrumental rationality that makes sophistication a plausible principle of instrumental rationality, namely *open prospects*, consequentialism turns out not to be a requirement of instrumental rationality. Is there a different way of thinking about the standard of instrumental rationality that would make consequentialism a plausible requirement of instrumental rationality? To avoid the problems we just pointed out, this would have to be a standard that is not shifting in the same way as *open prospects* was.

Consequentialism would be a requirement of instrumental rationality if the agent's preferences over the prospects available to her at the outset of the decision problem remained the standard against which her later choices are judged. As we already noted, the agent in fact retains those preferences throughout. According to *open prospects*, those preferences are no longer instrumentally relevant once some uncertainty has been resolved. They may remain relevant, however, if the standard of instrumental rationality was that we should choose such that we bring about our most preferred prospect out of the prospects initially open to us. Call this *initial prospects*.

Note that *initial prospects* does not require the agent to act by her initial preferences over the prospect she faces initially. This would hardly be defensible as a notion of instrumental rationality, since it would take something other than the agent's current conative attitudes to be the standard of instrumental rationality. Instead, it requires her to act well according to her current preferences over the initial prospects the actions currently open to her would form part of. According to *initial prospects*, for decisions that happen after some uncertainty has been resolved, this bygone uncertainty remains relevant for what the agent should choose. In our example, instrumental rationality would now require the agent to do well by her Allais preferences, even after the uncertainty has been partially resolved at t_2.

Initial prospects seems to offer an instrumental justification for consequentialism. The preferences over the prospects initially open to the agent form a stable standard against which to evaluate the agent's actions. And an agent who violates consequentialism clearly does worse by her preferences over the prospects initially open to her. However, *initial prospects* also implies that sophistication is not a requirement of instrumental rationality.

The example of Rory brings this out. By choosing to go 'down' and play it safe at t_2, Rory acts in a sophisticated manner. But this choice is not endorsed by *initial prospects.* To do well by his preferences over the initial prospects open to him, he would have to choose 'up'. Sophistication seems instrumentally rational when the agent's preferences over the prospects open to him at the time of action are the standard of instrumental rationality. But if preferences over prospects including bygone uncertainty are the standard of instrumental rationality, the agent may be required to violate sophistication.

Hence, *open prospects* can justify sophistication, and *initial prospects* can justify consequentialism, but neither account of the standard of instrumental rationality can justify both principles. And then neither *open prospects* nor *initial prospects* condemn an agent with Rory's preferences as irrational. Agents with Rory's preferences can at most abide by one of sophistication and consequentialism. But according to each standard, that would be enough. According to *open prospects*, there is nothing wrong with Rory if he is sophisticated and violates consequentialism. And according to *initial prospects*, there is nothing wrong with Rory if he violates sophistication and abides by consequentialism.

Of course, if Rory adopted separable preferences he would abide by both requirements at the same time. The problem, however, is that we cannot explain what the instrumental appeal is of abiding by both principles, since they only seem attractive on distinct notions of what the standard of instrumental rationality is. And so an agent who doesn't already have separable preferences can't be given an instrumental reason to abide by both, and adopt separable preferences.

One last proposal may be that instrumental rationality requires the agent to do well by both her preferences over the initial prospects available to her as well as by her preferences over the prospects still open to her. As we have seen, if she has Allais preferences, then these two sets of preferences will be in conflict in the example we considered. However, that still does not seem to give her reason to change her preferences such that this conflict disappears. It merely means that she should find some compromise between conflicting sets of preferences.

I therefore think that the proof presented by Hammond and reconstructed by McClennen should do nothing to convince us that agents like Rory are instrumentally irrational. We can justify at most one of the two crucial rationality requirements that Hammond presupposes. We can justify consequentialism under *initial prospects*, and we can justify sophistication under *open prospects.*

In the following, I want to argue that both of these conceptions of the standard of instrumental rationality in fact still share a common flaw. And that is that both proposals assume that it is preferences over prospects — be it preferences over prospects including, or not including bygone risks — that form the standard of instrumental rationality. The next section argues that if we allow such preferences to form part of the standard of instrumental rationality, we cannot in fact justify any general principles of choice under uncertainty instrumentally. In particular, the best case for what is in fact instrumentally irrational about Rory's choice behaviour can only be made once we abandon this assumption.

5.6 Prospects and the Standard of Instrumental Rationality

As we have seen in Chapter 3, in the context of certainty, there is a question as to whether preferences over outcomes, or desires over simpler states of affairs that are features of those outcomes are the standard of instrumental rationality. I have argued that it needs to be the latter if there should be any hope for money pump arguments in favour of transitivity to go through. In the context of uncertainty, there is an additional question we need to answer. The question is whether conative attitudes that apply directly to prospects rather than merely to outcomes are permitted as the standard of instrumental rationality.

As we saw in the last section, the rationality principles presupposed in Hammond's dynamic argument for expected utility theory only seem like straightforward principles of instrumental rationality if preferences over prospects directly serve as the standard of instrumental rationality. In a sense, this is a natural extension of *preference-based instrumental rationality.* In the context of certainty, outcomes are the object of choice. *Preference-based instrumental rationality* takes preferences over those objects of choice to be the standard of instrumental rationality. In the context of uncertainty, prospects are the objects of choice. The positions we just considered take preferences over those to be the standard of instrumental rationality. I here want to argue that this cannot be so if we want there to be any hope of justifying even the most uncontroversial requirement of rationality in the context of uncertainty.

The requirement of state-wise dominance is much less controversial than separability, and accepted even by most rivals of expected utility theory.[8] It is in fact implied by the sure-thing principle and independence.

State-Wise Dominance: For any two actions A_i and A_j, if for every state of the world S_i, $O_{ii} \succcurlyeq O_{ji}$, then $A_i \succcurlyeq A_j$. If, in addition, $O_{ii} \succ O_{ji}$ for at least one state of the world S_i, then $A_i \succ A_j$.

This requirement states that if an action leads to a weakly preferred outcome than another available action in every state of the world, then that action ought to be weakly preferred. Moreover, if it is strictly preferred in at least one state of the world, then the action ought to be strictly preferred. For instance, in the cycling example presented in Section 5.2, if it were the case that I actually prefer taking the subway to cycling whether it rains or not, state-wise dominance would require me to prefer to take the subway.

State-wise dominance seems like a fairly uncontroversial requirement of instrumental rationality. At least it does so if we take the agent's preferences over outcomes to be the standard of instrumental rationality, and assume *maximization*: A maximizing agent who violates state-wise dominance seems to do worse by her preferences over outcomes no matter what happens. However, we have already rejected *preference-based instrumental rationality* in favour of a desire-based notion of rationality. This may make us worried about state-wise dominance.

First, given a desire-based notion of instrumental rationality, we may be sceptical of *maximization*, especially if preferences are understood to themselves be conative attitudes. But in fact, we can simply formulate state-wise dominance as a principle about actual choice. This circumvents worries about *maximization.* And presumably the point even behind the original version of state-wise dominance was that an agent who *chooses* such that she does worse no matter what seems to be instrumentally irrational.

Second, we may be worried about state-wise dominance under a desire-based notion of instrumental rationality because we can imagine cases of non-uniqueness where it would be consistent with desire-based instrumental rationality to violate state-wise dominance. Take again the cycling case. Suppose that I prefer taking the subway to cycling no matter whether it rains or not. But suppose that it would have also been permissible to have the preferences described in Section 5.2, whereby I prefer cycling to taking the subway if it is sunny, but have the opposite preference if not. Acting in accordance with preferences that I do not have, but that would have also captured my underlying desires well seems

[8] I follow McClennen's (1990) formulation of what he calls 'dominance in terms of sure (riskless) outcomes', or DSO (p.50). See also Buchak (2013), p.94, who requires state-wise dominance in her rival theory to expected utility theory, just as Quiggin (1982) does.

permissible under desire-based instrumental rationality. In this case, this may involve cycling, and thus violating state-wise dominance.

Still, the following version of state-wise dominance gets around this problem:

State-Wise Dominance*: An agent ought not to choose an action A_i if there would have been an alternative action A_j, such that for every state of the world S_i, O_{ii} is just the same as O_{ji}, except that in at least one state of the world S_i, O_{ii} has more of something the agent desires[9] than O_{ji}.

Violating this requirement would mean doing worse by your desires no matter what happens. Consider, for instance, a case where somebody offers to give you a free coffee on your subway ride in the event it starts to rain later. If you refuse it, your subway ride will be the same whether it rains or not. And suppose that accepting the coffee would change nothing but satisfy your desire for coffee in the event that it rains later. Refusing the offer would then be in violation of state-wise dominance*.

State-wise dominance* in fact only seems to be an instance of the following more general, but vaguer, state-wise dominance**:

State-Wise Dominance:** An agent ought not to choose an action A_i if there would have been an alternative action A_j, such that for every state of the world S_i, O_{ii} is just the same as O_{ji}, except that in at least one state of the world S_i, O_{ii} is definitely favoured over O_{ji} by the agent's desires.

Both of these principles are meant to capture that it is irrational to make a 'sure loss' with regard to one's desires over outcomes. Are these versions of state-wise dominance, then, uncontroversial requirements of instrumental rationality? They are so only if the desires that form the ultimate standard of instrumental rationality apply to features of outcomes only, and not also to (features of) prospects directly. Suppose, for instance, that I have a strong desire for secure prospects. This desire is satisfied whenever I choose a prospect that leads to the same outcome in every state of the world. If I have such a desire, and that desire is strong enough, desire-based instrumental rationality does not seem to prohibit me from violating state-wise dominance*.

Take our coffee-on-the-subway example again. If I have a strong desire for secure prospects, I may refuse the offer of free coffee in the event of rain, because that offer would result in an uncertain prospect. My dislike of such prospects outweighs my desire for coffee. Sure, my desire to have coffee pulls me the other way. For this reason, we cannot understand the desire for secure prospects to be an instrumental desire that ultimately aims at something else. But we are allowing my desire for secure prospects to be part of the standard of instrumental rationality, and thus be a non-instrumental desire in its own right. And my desire for certain prospects speaks in favour of refusing the offer of coffee in the event of rain. If that desire is strong enough, why shouldn't it be instrumentally rational to refuse the offer? Instrumental rationality cannot require me not to have this kind of desire.

Hence, even a seemingly uncontroversial principle of instrumental rationality like state-wise dominance* cannot be a general principle of instrumental rationality if we allow desires regarding prospects directly to be part of the standard of instrumental rationality. In fact, if we admit these desires, it seems like we cannot formulate any general principles about how our preferences over prospects should relate to our preferences or desires over outcomes. Whatever those proposed principles are, we can imagine an agent who has a strong desire for prospects the choice of which would violate the principle. If instrumental rationality can't forbid the agent to have that desire, then it cannot justify the principle.

[9] Again, as in principle Q (Section 3.7), I mean these objects of desire to be features of outcomes that are individuated finely enough such that they themselves are not desired by the agent in one respect, while she is averse to them in another.

Critics of expected utility theory often point out that agents do seem to be sensitive to features of prospects directly. For instance, Lopes (1981, 1996) argues that next to certainty, mean, mode, variance, skewness and probability of loss are further 'global' features of gambles agents may care about. Buchak (2013) calls agents who are sensitive to values that are only achieved though a combination of outcomes across different states (other than expected utility itself) 'globally sensitive'. These critics argue that expected utility theory cannot account for this sensitivity.

However, if we were to take this sensitivity to imply that agents have desires for prospects that are not merely instrumental, and those desires form part of the standard of instrumental rationality in their own right, then, as we have seen, we cannot justify any general principles on how preferences over prospects ought to relate to desires or preferences over outcomes. We cannot even justify state-wise dominance*. But state-wise dominance, even in its original form, is accepted by most critics of expected utility theory.

We may respond to this by restricting the scope of the requirements of decision theory to those agents who do not have non-instrumental desires for prospects, or at least not non-instrumental desires for prospects that may speak in favour of violating one's favourite principles of choice under uncertainty. However, this may be giving up too quickly. Decision theory was intended to be a general theory of instrumental rationality. Restricting its scope to only agents who have or don't have a certain kind of desire should be a last resort — as it was when we reached our conclusion in Chapter 4.

An alternative response could be to reinterpret any desire that seems to be directly and non-instrumentally about features of prospects to be either reducible to a desire about outcomes, or to be an expression of a sensitivity to risk that is compatible with desires for prospects being merely instrumental. The first, indeed, seems to be the orthodox response.

Reducing desires regarding features of prospects to desires about outcomes may work as follows in our example: That it is part of a certain prospect also seems to be a feature of the outcome of taking the subway having refused the coffee, and that it was part of an uncertain prospect is likewise a feature of the outcome of drinking coffee on the subway. My desire for certain prospects may be fully accounted for by my desires over outcomes thus described. And then we can justify state-wise dominance* instrumentally after all, as well as perhaps other principles, such as separability. As long as I prefer taking the subway without coffee as part of a certain prospect to taking the subway with coffee as part of an uncertain prospect, I do not violate state-wise dominance* by refusing the offer of coffee in the event of rain.

Indeed, it is commonly held that reasons for action must ultimately derive from what things will be like in some state of the world. Broome (1991) appeals to such a claim in his defence of separability. Buchak (2013) calls the claim that a prospect can be preferred over another only if it is better in some state 'betterness-for-reasons' (p.75), and appeals to it in order to justify her version of state-wise dominance. She thus also takes reasons for action to derive from our evaluations of outcomes in states. In the case of instrumental rationality, where our reasons for action derive from our desires, the claim is that our desires, or at least the desires relevant for instrumental rationality, only concern outcomes.

I will, in the following, adopt this common assumption. First, this is because the assumption is more innocuous than it may seem. Nothing stops us from including the precise structure of the prospect an outcome is part of in the description of an outcome. Buchak (2013) calls this 'global individuation'.[10]

[10] Global individuation is also often appealed to in order to defend expected utility theory against counterexamples. For instance, Weirich (1986) argues that globally sensitive aversion to risk can be represented with disutilities that are assigned to outcomes. In the context of Buchak's theory, Pettigrew (2015) argues that the global sensitivity allowed for by her theory is compatible with expected utility theory if outcomes are appropriately redescribed. In particular, he proposes

That way, desires that are about prospects can be translated into desires that are about outcomes.[11]

Second, even if this assumption is not innocuous,[12] and 'betterness-for-reasons' is not as intuitive as Broome or Buchak suppose, the assumption is still necessary for there to be any chance of justifying principles of standard decision theory under uncertainty instrumentally. As we have seen, even something as weak as state-wise dominance* cannot be justified as a general principle of instrumental rationality unless we think the agent's non-instrumental desires are only about outcomes. And so even if the assumption is suspect, we will accept it for the sake of argument. It is indeed accepted by defenders and critics of expected utility theory alike. Not accepting it would mean already giving up on the idea that general principles of decision-making in the context of uncertainty could be instrumentally justified independently of what desires an agent may have.

The assumption that the desires that form the standard of instrumental rationality are desires for features of outcomes gives us a straightforward justification for state-wise dominance*. An agent who violates it will do worse by her desires over outcomes no matter what happens. However, it is not obvious that this standard of instrumental rationality can help us justify any further requirements on preferences over prospects.

Once the standard of instrumental rationality is desires over features of outcomes, it initially seems like we can be fairly permissive of different preferences over prospects given the same desires over outcomes. Suppose I desire only muffins. I then consider the question of whether I would be willing to forego 40 muffins for the chance to win 100 muffins in a fair coin toss. It appears like either answer is compatible with me being instrumentally rational. And that is because it is not clear which option serves my desire for muffins better. Sure, we may not allow just any preference over prospects that does not violate state-wise independence* to be instrumentally rational. For instance, in normal circumstances, it may not be instrumentally rational to forego 99 muffins for the chance to win 100 in a fair coin toss. But as long as our preferences over prospects are not this extreme, why shouldn't instrumental rationality allow for them, given only desires over outcomes are the standard of instrumental rationality?

It does not help here to invoke strength of desire, or even a cardinal utility function hat measures strength of desire. As we said above, it seems like agents may have different preferences over uncertain prospects, even given the same cardinal utility function measuring strength of desire for outcomes. In our

that the gamble by which the outcome was achieved should be part of the description of the outcome. While this shows again that 'betterness-for-reasons' is widely accepted, this move only helps expected utility theory if we can give some positive defence of separability as a principle of instrumental rationality. The following shows that this is harder than it may seem.

[11] As Buchak points out, the problem with this strategy is that it leads to a proliferation of outcomes that ultimately makes it difficult to give any structure at all to choice under uncertainty (see p.139-45). Another problem is that the completeness assumption that most decision theories make requires agents to have preferences over all possible distributions of outcomes over states. If we individuate globally, this means that agents also have to have preferences involving a prospect that has me take the subway as part of a certain prospect, if my coin lands heads, and cycle in the sun as part of a certain prospect, if my coin lands tails. Such prospects are not possible objects of choice, which is why Broome (1991) calls preferences over these kinds of prospects 'impractical preferences'. However, it is not clear that standard decision theory can do without presupposing such impractical preferences. It may also help to note that, as we mentioned in Chapter 2, most decision theorists think that completeness is not a requirement of rationality, but that instead coherent extendibility is enough. Perhaps, then, agents need not have these impractical preferences after all. If agents do have them, however, prospects such as the one I just described may also make us sceptical that global individuation can really capture desires over features of prospects adequately. If everything there is to be captured about my love for certain prospects is captured by my desires over globally individuated outcomes, and if I prefer cycling in the sun for certain to taking the subway for certain, then I should prefer the impossible prospect just described to the prospect of taking the subway for certain, no matter what the outcome of the coin toss is. But for agents who desire certain prospects, this may not be plausible.

[12] For further arguments against this assumption, see Stefansson and Bradley (2015, forthcoming). They instead defend the view that chances can have non-instrumental value, resulting in their value being non-linear in probabilities. My worry is that once we allow for chances to have non-instrumental value, we can, as Humeans about decision theory, no longer require that the value of chances should even be increasing in probabilities. And then not even state-wise dominance seems justifiable.

example, suppose that I desire each muffin the same, no matter how many muffins I already have. This can be expressed in a linear utility function over muffins. It still seems like, given such a utility function, both accepting or refusing the coin toss would be permissible in so far as instrumental rationality is concerned.

The plausibility of such permissiveness is one of the main motivations for Buchak's (2013) alternative to expected utility theory. She argues that due to this permissiveness, it can be perfectly rational for agents to be sensitive to some global features of prospects. Buchak (2013) argues for permissiveness to global sensitivity within bounds — notably the bound of state-wise dominance. In particular, she thinks that it can be rational to weight what happens in the worst states disproportionately heavily, resulting in an aversion to taking risks. The next chapter will discuss her theory in more detail.

For now, note that the sensitivity Buchak has in mind here is not explained by appeal to some non-instrumental desire for a certain kind of prospect. The agent views prospects as instrumental to fulfilling her desires for outcomes. Buchak merely argues for permissiveness when it comes to how the agent structures her attainment of the desired outcomes. Nothing we have said so far speaks against instrumental rationality being permissive in this sense. And this is, thus, the second way, next to global individuation, in which we can account for global sensitivity without admitting attitudes over prospects to be themselves the standard for instrumental rationality.

In the following, I will take permissiveness about attitudes over prospects to be the default position given desires over features of outcomes are the standard of instrumental rationality. The burden of proof lies with those who want to justify requirements on preferences over prospects that go beyond state-wise dominance*, and perhaps the requirement that agents shouldn't take risks that are too great. Given this default position, separability can at least not be justified in any straightforward way. The Allais case in fact brings this out.

Consider again the second Allais choice. Rory's preferences are such that he would strictly prefer D if one of tickets 90-100 is drawn, and he is indifferent when he knows that one of tickets 1-89 is drawn. Still, he prefers C when he does not know anything about which ticket has been drawn. He thus violates separability. But now consider that, given the standard of instrumental rationality is just the agent's desire for money, both choosing C and choosing D seems permissible when choosing between the complete prospects. Likewise, even given one knows that one of tickets 90-100 has been drawn, both choices seem like permissible ways of trying to satisfy one's desire for money. But then given such permissiveness, we cannot say that there is anything obviously wrong with having non-separable preferences. Each of Rory's preferences seems permissible given his desire for money. And whichever way he chooses, there is a good chance that he would have gotten more money had he chosen differently.

	Tickets 1 - 89	Tickets 90 - 99	Ticket 100
Lottery C	$0	$5 million	$0
Lottery D	$0	$1 million	$1 million

Table 5.5: Allais Paradox: Second Choice

Separability is not the only principle that cannot be justified in a straightforward way once we take only the agent's desires over outcomes to be the standard of instrumental rationality. In fact neither consequentialism nor sophistication come out as straightforward principles of rationality anymore given this account of instrumental rationality.

As we saw in Chapters 2 – 4, *maximization* cannot be taken for granted if preferences are understood to be conative attitudes, but instrumental rationality is desire-based. In particular, if there is non-uniqueness in what preferences are permissible given the agent's desires, then *maximization* is no longer required by instrumental rationality. If instrumental rationality is permissive about how agents choose, or which preferences they may have between prospects given their desires over outcomes, then the very same applies here. Instrumental rationality cannot require agents to choose in accordance with their preferences over prospects — be it the prospects open to them at the time of decision, or the initial prospects open to them. As long as the agent ends up choosing some prospect that is permissible given her desires over outcomes, then instrumental rationality cannot criticize her. But sophistication and consequentialism rely on such requirements to be guided by one's preferences. Neither sophistication nor consequentialism then seem defensible.

This assumes, however, that agents *can* act against their preferences. As was the case with *maximization* in Chapter 4, it may help to reinterpret preferences as dispositions to choose. Most plausibly, we could understand the agent's preferences over the prospects open to her at the time of action as her binary dispositions to choose between any two of them. In that case, if the agent only makes binary choices, sophistication follows given that the agent makes accurate predictions of how she will act in the future.

However, note that, as before, sophistication when more than two options are involved at any one point still needs defence. Moreover, this does not guarantee us sophistication with regard to the agent's *given* preferences, as is needed for Hammond's proof — agents can always adjust their preferences momentarily, as in one form of resolution. In fact, as we will see below, the possibility of resolution undermines the best argument in favour of separability. Lastly, consequentialism cannot be defended in this way, since it may require the agent to take bygone risks into consideration and act against her preferences over the prospects open to her at the time of action.

We have found, thus, that if we want to justify even the most uncontroversial principle of choice under uncertainty instrumentally, we have to allow only for desires regarding outcomes to form part of the standard of instrumental rationality, rather than also for attitudes over prospects. But if that is so, separability does not come out as a straightforward principle of instrumental rationality. And neither do the two principles Hammond uses to derive a requirement to have separable preferences.

We are hence in need of a different defence of separability. In the following, I want to argue that a better version of the dynamic choice argument in favour of separability points out that agents with separable preferences may end up violating state-wise dominance* *over time.*

5.7 Dynamic Dominance Violations

It can be shown that an agent like Rory may end up choosing a course of action that leaves him with a worse outcome, no matter what happens, than another course of action he could have engaged in — even though no individual choice he makes is a sure loss choice. Suppose Rory is offered the opportunity to pay some small cost ϵ,[13] at the beginning of the decision problem, in order to bind himself to the choice he prefers at the outset. This alters the second decision problem to the one pictured in Figure 5.2, where t_0 is the point in time at which he can bind himself to lottery C at cost ϵ. As a sophisticated

[13] As in previous chapters, the costs involved in this sure loss case need not be monetary. Perhaps pre-commitment involves a social interaction that Rory is anxious about. Perhaps it involves wasting some precious ink.

agent, he should choose to in fact bind himself in this way. He knows that if he does not do so, he will choose in accordance with lottery D. If he has a strict preference, at the outset, for C over D, there will be a small enough ϵ such that he prefers to go 'down' at t_0 and bind himself.

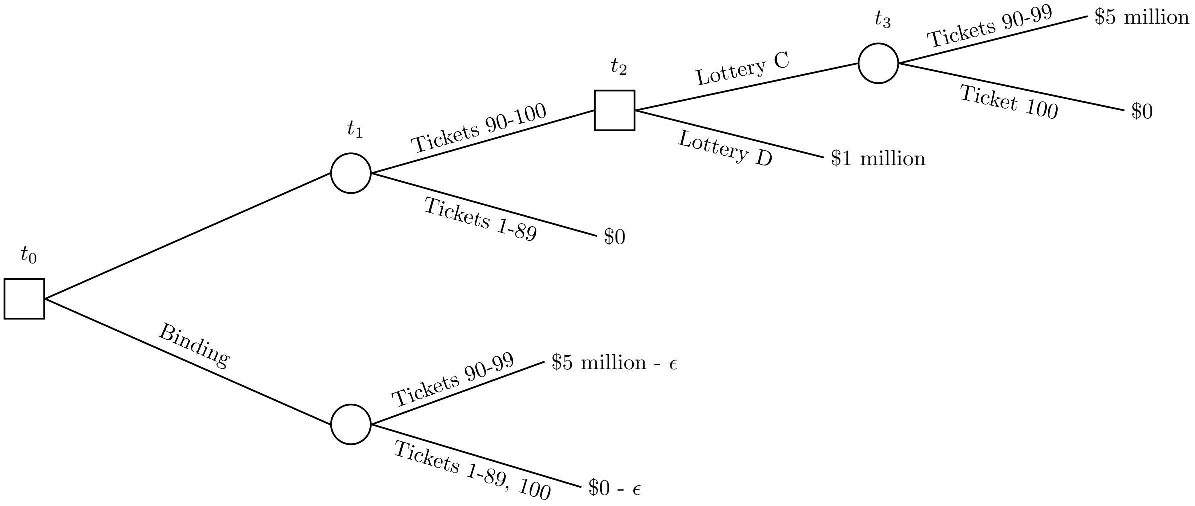


Figure 5.2: Alternative Second Choice

This result is worrying because no matter what happens, Rory will end up with an outcome he dis-prefers to the outcome he would have had, had he taken another course of action that was available to him.[14] The course of action where he chooses not to bind himself, and then chooses to go 'up' at t_2 is also available to him. In that case, he also ends up choosing in accordance with lottery C, but he avoids paying ϵ.[15]

Note also that not only does he end up with a dis-preferred outcome no matter what happens, he also ends up with an outcome that is the same as the one he could have had, except that it has less of something he desires. Rory thus ends up, over time, violating state-wise dominance*, if we understand state-wise dominance* as a principle about entire courses of action. His course of action over time then turns out to be criticizable according to desire-based instrumental rationality, given that desires over outcomes only are the standard of instrumental rationality. This is a stronger result than the one we noted in the previous dynamic choice problem. It is not only that the agent ends up with a prospect that he dis-prefers at the outset of the decision problem. The course of action he chooses is in fact state-wise dominated* by another course of action that was available to him: He ends up with a worse *outcome* no

[14] Another reason why this result has been held to be worrying is that it seems odd that an agent who expects to be rational in the future, and expects to have the same preferences over outcomes in the future, would choose to restrict his own future freedom of choice. One could support this thought by arguing that it is in most agents' interest to retain as much autonomy as they can. Binding oneself to some future course of action seems to be called for in situations where one faces temptations or impairments of one's future judgement. But neither of these conditions hold here. In fact, defenders of Allais preferences want to maintain that the complete set of Allais preferences is justified.

[15] Another argument why Allais preferences may lead to unacceptable choices in dynamic choice problems is based on a similar variation of the decision problem. Suppose our sophisticated agent is offered the choice to not receive the information about whether one of tickets 90-100 have been drawn at t_1. If she faces the second choice, and anticipates to go 'down' at t_2 if she is given the information, then as a sophisticated agent she should refuse the information. Without the information, she will simply face the original, one-off Allais problem, and choose lottery C. If she gets the information, as a sophisticated agent, she will end up choosing in accordance with lottery D. Knowing this, before any uncertainty has been resolved, the agent should refuse to get this information, even if it is free. In fact, she should be willing to pay some small amount of money in order not to receive the information. Refusing free information has also been argued to be irrational in its own right. See, for instance, Wakker (1988).

matter what happens.

There is a parallel here to the temptation case we considered in Chapter 2. In those cases, too, we were worried about the agent engaging in costly pre-commitment, when it would have been possible to get the benefits without those costs. The same holds here: The agent has a definite interest in avoiding costly pre-commitment, since that would leave him with a sure loss.

We may be worried about requiring the agent's entire course of action to abide by a version of state-wise dominance*. We may, for instance, imagine cases where the agent's underlying desires change in such a way that abiding by a diachronic version of state-wise dominance* would require the agent, at some point, to make a choice that is not endorsed by the agent's desires over the outcomes still available at the time of action. For the same reason, we were worried about the diachronic part of Andreou's principle P in Chapter 3.

However, this weaker version of the diachronic requirement for the agent's courses of action to abide by state-wise dominance is not subject to that worry:

> S: It is irrational to make a series of choices C_1, when
>
> 1. there would have been an alternative series C_2, such that for every state of the world S_i, O_{2i} is just the same as O_{1i}, except that in at least one state of the world S_i, O_{2i} has more of at least one thing the agent desires than O_{1i}.
> 2. there is an alternative available series of choices of which (1) is not true, and whereby each individual choice would have been supported by one's desires concerning the outcomes still available at the time of action.

Note that this principle is parallel to principle Q introduced in Section 3.7. In fact, Q turns out to be just a special case of S, in case there is only one state of the world. Clause (2) expresses the thought that sure loss over time may not be irrational when the only ways to avoid it would involve a choice at some point in the decision problem that is itself instrumentally irrational — as it may be the case with the standard temptation cases considered in Chapter 2.

Likewise, we can formulate a diachronic version of the more general state-wise dominance**:

> T: It is irrational to make a series of choices C_1, when
>
> 1. there would have been an alternative series C_2, such that for every state of the world S_i, O_{2i} is just the same as O_{1i}, except that in at least one state of the world S_i, O_{2i} is definitely favoured over O_{1i} by the agent's desires.
> 2. there is an alternative available series of choices of which (1) is not true, and whereby each individual choice would have been supported by one's desires concerning the outcomes still available at the time of action.

This principle, in turn, is parallel to principle R introduced in Section 3.7. Both S and T express the idea that it is irrational to make a 'sure loss' over time with regard to one's desires over features of outcomes (provided avoiding to do so does not itself require an instrumentally irrational choice).

In the dynamic choice problem just presented, Rory appears to violate S. We already saw that he chooses a dominated course of action. Moreover, avoiding taking a dominated course of action does not seem to require him to engage in any actions that are themselves problematic in terms of instrumental rationality, at least not if we are permissive in the way we argued for in the last section.

One thing that may be said against this assessment of what is instrumentally irrational about Rory's choices is that the alternative course of action whereby he gets lottery C without costs of commitment is in some sense not available to Rory. Given the fact that he is sophisticated with regard to his given preferences over prospects, it is not open to him to choose to go 'up' at t_2. If this course of action is not available to Rory, then Allais preferences will not lead the sophisticated agent to choose a course of action that is dominated by another one that would have been available to him. This is the line that Seidenfeld (1988) takes in response to this argument.

Along with McClennen (1990) and Steele (2010), I do not find this response satisfying. For one, it is up to Rory himself that the course of action whereby he ends up with lottery C without paying ϵ is not available to him. No external factor is keeping him from taking that course of action. In fact, it is because he is sophisticated with regard to his given preferences that this course of action is not available to him. However, sophistication is itself in need of instrumental defence. Sophistication, on the face of it, seems like an intuitive requirement of instrumental rationality. However, as we have seen, it is only defensible as such a requirement on an understanding of the standard of instrumental rationality that we have abandoned — namely on an understanding that takes attitudes over prospects directly to be the standard of instrumental rationality. If only desires over outcomes are the standard of instrumental rationality, an agent does not seem to be required to be sophisticated. And so Seidenfeld's response does not work.

I therefore think that this is a convincing argument showing that something is wrong with Rory: He chooses a course of action over time that is strictly worse with respect to his desires over outcomes than another available course of action. And he could have avoided this without acting in a way that is itself instrumentally criticizable. His course of action over time is thus clearly instrumentally deficient.

5.8 Resolution

We might think that the fact that Rory violates S gives him a reason to adopt separable preferences. Chapters 3 and 4 considered an instrumentalist argument in favour of the transitivity of preference that was in many ways similar to an argument to that effect. The money pump arguments we considered there are arguments that agents with intransitive preferences who act in accordance with those preferences will, over time, do badly with respect to their desires over outcomes. In particular, they will violate principle Q, which is just the special case of S when there is only one possible state.

As we saw, the problem with those arguments is that an agent need not adopt transitive preferences to avoid being money pumped. She may also either fail to act in accordance with her preferences at some crucial points in time, or merely adopt a temporary preference to do otherwise than she normally would. Being resolute in the way McClennen (1990) describes it, or acting in a unified way in the way Rabinowicz (2014) describes it would be effective ways of doing so. We argued that those ways of deciding are not instrumentally irrational, since we cannot give an instrumental justification for the requirement to act in accordance with our preferences, or to have stable preferences over time and across different choice situations, on the most plausible account of the standard of instrumental rationality.

The very same problem arises here. We have argued that Rory's choices over time are not instrumentally rational. Rory has Allais preferences throughout, which violate separability, and behaves in a sophisticated way. He could either avoid instrumental irrationality by adopting stable separable preferences and remaining sophisticated. Or he could avoid it by failing to be sophisticated with regard to his given preferences. As long as this alternative response is available to him, it seems like we cannot justify the requirement to have separable preferences instrumentally.

How could an agent avoid being instrumentally irrational over time by failing to be sophisticated with regard to his given preferences? In our example, Rory could choose, at t_2, to go 'up', against his given preference for the safe prospect available at that point in time. Moreover, he would have to anticipate that he would do that at t_0, and thus choose to forego paying ϵ to bind himself. McClennen's resolute agents, and Rabinowicz's unified agents, for instance, would choose in this way.

As we saw in Chapter 3, agents who are resolute in the way McClennen describes them make a plan at the beginning of a dynamic choice problem to act in a particular way and then simply go through with it. In Rory's alternative decision problem, if he were a resolute agent, he would plan just before t_0 to go 'up' at t_0, and then 'up' again at t_2. When he gets to t_2, he would actually go through with that plan. And so the resolute agent does not choose the dominated course of action. He ends up with lottery D without paying ϵ. Again, Rabinowicz' unified agents behave in the same way. Unified agents consider at every point what course of action they would then choose for the decision problem as a whole, and act in accordance with that course of action.

As we have also seen before, there is a question as to whether the resolute agent is best modelled as acting counter-preferentially or as changing her preferences in the context of the dynamic choice problem, to conform to what she has planned to do at the outset. While McClennen (1990) and Machina (1989) prefer the latter strategy, Gauthier (1994) chooses the former. If we choose the former strategy, resolute Rory can't be sophisticated. At t_2, he chooses to go 'up' even though his preferences over the prospects available to him at that point in time favour going 'down'. If we choose the latter strategy, however, Rory in fact remains a sophisticated agent with regard to the new and altered preferences. He just can't be sophisticated with regard to his given preferences. Instead, he changes his preferences within the dynamic choice problem such that the course of action that resolution recommends is favoured by his preferences over the prospects still available to him at each point in time.

In either case, the agent need not adopt fully separable preferences. In one case, Rory just keeps his Allais preferences, and in the other case, he alters them only within the dynamic decision problem at hand. Is there anything that instrumental rationality can say against adopting one of these alternative responses to the dynamic argument presented in the last section?

As we have seen, many authors have been sceptical of the claim that counter-preferential choice can be instrumentally rational. I take this to be motivated by the idea that the agent's preferences define her ends in action, and are themselves the standard of instrumental rationality. In this case, the preferences against which resolute agents act are preferences over prospects. But we have already argued that we should take only the agent's desires over outcomes to be the standard of instrumental rationality, if we want there to be any hope of justifying principles of choice under uncertainty instrumentally.

As we argued before, if the agent's preferences over prospects are not themselves part of the standard of instrumental rationality, there seems to be no special reason why we shouldn't be permissive with regard to how agents may choose between different prospects. And then there seems to be no reason to suppose that agents ought to act in accordance with their preferences over prospects (provided they can). Moreover, those who want to argue that the agent ought to adopt separable preferences must think that at least one separable preference relation over prospects would do justice to the agent's desires over outcomes. As long as the resolute agent acts in accordance with those separable preferences within the dynamic choice problem, it thus does not seem like we can accuse her of being instrumentally irrational.

Lastly, resolution in its second guise does not require counter-preferential choice, and so worries about counter-preferential choice do not apply there.

We have hence reached the same impasse that we reached with the agent who has cyclical preferences. Rory's choices are certainly instrumentally irrational. But to do better, he need not adopt separable preferences — as long as he either acts counter-preferentially, or adjusts his preferences temporarily at the right point in time to avoid violating principle S.[16] And so we cannot offer a general instrumental justification for separability as a principle of instrumental rationality.

5.9 Conclusions

The core requirement of standard decision theory in the context of uncertainty is separability. We have argued here that just like the core principles of standard decision theory under certainty, separability cannot be justified as a general principle of instrumental rationality. In Chapter 4, we however also argued that the core principles of standard decision theory in the context of certainty could be defended to agents who have a desire to have stable preferences over time and across different choice situations and contexts.

The very same desire would help us here, too. If preferences are understood to be choice dispositions, and the agent has a desire for those to be stable in this way, then adjusting her preferences within particular decision problems only will frustrate that desire, and it will frustrate it unnecessarily. The agent could equally just adopt stable and separable preferences. For an agent who has such a desire, then, instrumental rationality seems to demand that she has separable preferences.

Still, separability cannot be defended instrumentally to agents who do not have such a desire. Expected utility theory thus turns out not to be a general theory of instrumental rationality. It is a theory of instrumental rationality only for those agents who desire to have stable choice dispositions. Agents who do not have such a desire are apparently free to violate separability.

Does that mean that no formal decision theory can provide us with an adequate theory of instrumental rationality? One advantage of having a formal decision theory is that it helps us give some structure to decision-making under uncertainty, or help us explain and predict an agent's behaviour under uncertainty. Perhaps an alternative to expected utility theory could still give us that, while being more permissive about the agent's preferences and choices in the context of uncertainty.

In fact, there are a number of alternatives to expected utility theory that relax the assumption of separability. The next chapter will turn to a recent and influential alternative, namely Buchak's (2013) REU theory. The challenge for any such theory is to offer a true alternative to expected utility theory, while providing us with a formal structure that does more than express the idea that instrumental rationality is permissive about our attitudes to prospects in the way we have described above. In particular, it should help us explain and predict the choices of instrumentally rational agents. The next

[16] One might think that being disposed to make these temporary adjustments, in the end, is not importantly different from having separable preferences. But note that making the kinds of temporary adjustments in question is consistent with having different preferences, and making different choices in situations where one is not about to be exploited. And it is consistent with adjusting one's non-separable preferences in different ways in different exploitable dynamic choice problems. It would thus undermine any decision theoretic analysis that models the agent as having a fixed separable preference relation. There is one caveat here. The distinction between having separable preferences and "temporarily" adjusting one's preferences within a dynamic choice problem does indeed seem to collapse if the agent treats her entire life as one long dynamic choice problem, as Section 6.6 also points out. As the next chapter argues, there is indeed a worry that instrumental rationality requires resolute agents to take such a long term perspective, thus undermining the resoluteness defence of non-separable preferences.

chapter argues that it is doubtful whether REU theory can do this.

Chapter 6

Risk Aversion and the Long Run

6.1 Introduction

If we want any chance of being able to justify the core requirements of standard decision theory instrumentally, then the standard of instrumental rationality must be desires over features of outcomes. The previous chapters have tried to establish this. One uncontroversial principle for choice under uncertainty that can be justified given this notion of instrumental rationality is state-wise dominance. It can't be rational to choose such that an alternative action would have left one with more of something one desires no matter what happens, or otherwise do definitely worse with regard to one's desires no matter what happens. That is, it is irrational to make a sure loss with regard to one's desires over outcomes. We indeed argued for diachronic versions of such a principle, namely principles S and T.

Beyond these principles, we said, it seems like we can be fairly permissive about the choices an agent makes in the face of uncertainty. If the agent's desires concerning features of outcomes are the standard of instrumental rationality, it does not seem like there could be anything irrational about an agent who avoids sure loss, as well as taking extreme risks, or foregoing almost certain benefits. We can try to show, however, that agents who violate further principles will in fact end up making a sure loss, and violating principles S or T. This is what the dynamic argument in the last chapter tried to do in the case of separability, the core requirement of expected utility theory.

As we showed there, however, separability can only be justified as a requirement of instrumental rationality to agents who desire to have stable preferences over time and across different choice situations. Separability thus ends up with the same status as the acyclicity of preference. In the case of both of these principles, agents who violate them can avoid the instrumental irrationality of making a sure loss in terms of their desires by either failing to act on their preferences over the options available to them at some point in time, or by adjusting their preferences within dynamic decision problems temporarily. In particular, resolute agents will respond to these dynamic arguments in one of these alternative ways. Appeal to resoluteness has in fact been a popular move in defending alternatives to expected utility theory.

There are a number of alternatives to expected utility theory that relax separability. Perhaps, then, there is some hope for alternative formal decision theories as theories of instrumental rationality. Such theories could potentially capture a more permissive picture of instrumental rationality while at the same time providing us with the benefits of having a formal decision theory to structure, explain and predict instrumentally rational decision-making under uncertainty. This chapter will focus, in particular, on Buchak's (2013) risk-weighted expected utility (REU) theory.

Apart from allowing the agent to avoid sure loss and abide by principles S and T, the success of any alternative formal decision theory seems to depend on firstly, offering a true alternative to expected utility theory, and secondly, helping to indeed structure, predict and explain instrumentally rational decision-making under uncertainty. If it does not do the latter, then the theory would appear to do little more than express the idea that instrumental rationality is permissive in the face of uncertainty.

Here I argue that REU theory struggles to do both. In particular, I will show that if we hold on to the assumption that agents are resolute, and moreover add the fairly uncontroversial requirement that agents formulate their decision problems such that everything relevant to their decision is included, then the agents REU theory describes end up behaving approximately like expected utility maximizers. And then REU theory cannot in fact account for the counter-examples to expected utility theory that motivated it in the first place. The same arguably holds for any theory that tries to account for our ordinary attitudes to risky prospects.

REU theory can avoid this conclusion by being more permissive about how agents formulate their decision problems, and by relaxing the assumption that agents need to be strictly resolute. In fact, both of these reactions seem to be permitted by desire-based instrumental rationality. But then the predictions of REU theory end up being extremely sensitive to the way in which agents choose to frame their decision problems, or choose to act in dynamic choice contexts. This calls into question whether the formalism of REU theory can capture any stable choice tendencies in the face of risk. And if it can't do so, it does not seem like it offers much beyond an expression of the kind of permissivism with regard to choice under uncertainty we just described.

6.2 Samuelson's Colleague

Critics of expected utility theory argue that expected utility theory does not offer a satisfactory treatment of risk aversion. In the last chapter, we already encountered one famous example in which a course of action that violates expected utility theory is both common and seems reasonable. As we also saw there, it is in fact difficult to justify the crucial requirement of separability, which these Allais preferences violate, as a principle of instrumental rationality. There are many more such cases, leading critics to argue that expected utility theory is too restrictive in what choices it allows in the context of uncertainty. This section will introduce another such case.

Most alternatives to expected utility theory have been introduced as descriptive theories of choice under uncertainty, with no claim to capturing rational choice. The most well-known is prospect theory, introduced by Kahnemann and Tversky (1979). Here, I focus on a recent alternative theory that was introduced explicitly as an alternative theory of instrumental rationality. And that is Buchak's (2013) REU theory. It is a rank-dependent theory similar to that found in Quiggin (1982). However, most of our conclusions will likewise apply for any alternative to expected utility theory that tries to make sense of the example discussed in the following.[1] This example features preferences that expected utility theory cannot easily make sense of, while Buchak's theory apparently can. It has accordingly served as a motivating example for the theory.

[1] For an overview of other alternatives to expected utility theory in the economic literature, the two most comprehensive surveys are Schmidt (2004) and Sugden (2004).

Samuelson (1963) reports having had the following lunch-time conversation with a colleague, who claimed to be a proponent of expected utility theory. Samuelson asked his colleague, henceforth SC, to consider a bet which gave him a 50% chance of winning $200 and a 50% chance of losing $100. SC refused on the grounds that this bet is too risky for him. However, he added that he would accept a series of one hundred such bets, since in the series of bets, he is almost certain to come out ahead. And indeed, the chance of making a loss on the series is below 0.5%, while the expected return is $5,000. The series of bets thus seems like an extremely attractive offer. At the same time, SC's preference to reject the individual gamble merely seems to display a reasonable level of risk aversion.

Samuelson himself shows that expected utility theory cannot easily make sense of these preferences, and thus seems to imply that SC is irrational. But, as I will show below, it appears that REU theory can accommodate SC's preferences. At least this is so when SC's decision problem is framed in the way it is by Samuelson. And so this seems like a reason in favour of REU theory: It can account better for ordinary risk averse preferences that seem intuitively rational.

The relevant outcomes in SC's case appear to be adequately described by the different amounts of total wealth SC may reach in the gambles he is offered. If we can assign a utility function to the agent, it will be a utility function over these wealth levels. We also assume that SC takes the probabilities he is given at face value, and that these also describe his degrees of belief. If he is an expected utility maximizer, he maximizes the sum of his utilities over wealth levels weighted by these probabilities.[2]

Given this, Samuelson shows that SC's preferences are incompatible with expected utility theory, under the plausible assumption that SC would prefer to reject each individual gamble at any wealth level that he might reach in the series of gambles (that is, at any wealth level between $9,900 below and $19,800 above his current wealth). In fact, for any gamble, and any number of repetitions, given this assumption, it is irrational to reject a single gamble that one would accept as part of a compound of many such gambles. What Samuelson shows is that in expected utility theory, under the assumptions he makes, gambles that are unfair in the utility metric, that is, have a negative expectation of utility, cannot be made fair by compounding them.[3]

I will not reproduce Samuelson's proof here. What is important for us is the feature of expected utility theory that is driving this and similar results.[4] And that is that in expected utility theory, risk averse behaviour, such as turning down a free bet with a positive expected gain in terms of money, is always explained by the concavity of the utility function in money. When a utility function is concave, the marginal utility derived from a good is decreasing: Any additional unit of the good adds less utility the more of the good the agent already has.

When the utility function is concave in this way, then setting the status quo at utility 0, the expected

[2] By treating probabilities as given in this chapter, we are following expected utility theory in the way it was developed in von Neumann and Morgenstern (1944), just like most economists.

[3] If we drop the assumption that the agent will reject the single gamble at all possible wealth levels that she could be at during a sequence of gambles, then, as Ross (1999) shows, it is possible for an expected utility maximizer to reject the single gamble and accept the compound. However, we might think that SC's preferences are sensible even if that condition is not met. If SC is sufficiently wealthy, then the differences in wealth in question here should not change his attitude to small stakes gambles. If we share this intuition, then Ross' result does not help us much in reconciling the pattern of preferences SC displays with expected utility theory.

[4] The case of SC illustrates a more general divergence between what seems reasonable and what expected utility maximization demands when it comes to large and small stakes gambles involving monetary pay-offs. Rabin and Thaler (2001) show that, for an expected utility maximizer, any significant risk aversion for small stakes gambles implies implausibly high levels of risk aversion for larger stakes gambles. Conversely, plausible levels of risk aversion for large stakes gambles imply near risk neutrality for small stakes gambles. For instance, an expected utility maximizer who turns down a 50/50 bet of losing $10 and winning $11 will turn down any 50/50 bet involving a loss of $100, no matter how large the potential gain. And Rabin (2000) shows that under expected utility maximization, SC should also reject a 50/50 gamble of losing $200 or winning $20,000.

utility of a gamble can be negative even if the expected money value of a gamble is positive. For the gamble in question here, this is so if the utility loss from losing \$100 is bigger than the utility gain from winning \$200. For instance, suppose the agent's utility in money is given by $u(m) = \sqrt{m}$, and the agent's current wealth is \$100. Then the expected utility of accepting SC's gamble is $0.5 * \sqrt{0} + 0.5 * \sqrt{300} \approx 8.66$. Rejecting the gamble yields $\sqrt{100} = 10$, which is higher. And so an agent with such a utility function should reject SC's gamble, and will thereby display risk aversion. Figure 6.1 illustrates this.

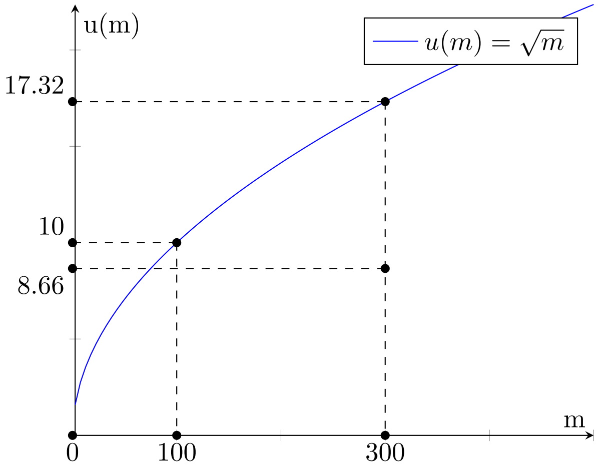


Figure 6.1: A Concave Utility Function

If we assume expected utility maximization, the risk aversion SC displays for the low stakes gamble must mean that the utility function we assign to him is concave in a way that makes a difference for small amounts of money. And I show here, drawing on Rabin and Thaler (2001), that this in turn implies extreme levels of risk aversion for larger stakes gambles, under the assumption that the agent would turn down SC's small stakes gamble at any relevant wealth levels. This extreme level of risk aversion should lead SC to reject the larger compound gamble which would guarantee him an almost certain gain with a high expected value.

Let current wealth be w. For SC to reject the small stakes gamble, the utility added by each dollar between w and $w+200$ can be at most $\frac{1}{2}$ of the utility added by each dollar between w and $w-100$. But given that we assumed that SC would also turn down the gamble if his initial wealth level was $w+200$, the added utility of each additional dollar between $w+200$ and $w+400$ can be at most $\frac{1}{2}$ of the added utility of dollar $w+200$, which is itself at most $\frac{1}{2}$ of that of dollar $w-100$. When we iterate this to larger sums of money, it turns out that dollar $w+2,000$ can add at most 0.2% of the utility of dollar $w-100$.

Hence, risk aversion for small amounts of money at wealth level w means that the utility gain by each additional dollar is only a fraction of what it is at w when SC is only \$2,000 richer. An extra dollar at $w+5,000$ adds virtually no utility compared to the utility gains at $w-100$. This is why a rejection of the gamble SC is offered implies extreme levels of risk aversion when it comes to larger stakes gambles, such as the compound gamble SC considers. Such an agent must be assigned a utility function with a rapidly decreasing marginal utility of money. If he rejects the small stakes gamble, SC should end up rejecting the large stakes gamble, because the larger sums of money he stands to gain from the larger stakes gamble add little utility for him.[5]

So if we want to insist that having preferences like SC's (and acting on them) is rational, but hold on to the assumption that the wealth levels that can be achieved in the gambles are the only relevant outcomes, then we have to abandon expected utility theory. In particular, we have to argue that the way expected utility theory treats risk aversion is inadequate: Risk aversion cannot be merely a matter of a concave utility function. And indeed, as we saw in the last chapter, expected utility theory, and in particular its treatment of risk aversion, has been controversial as a theory of practical rationality for other reasons. Ordinary decision-makers appear to violate expected utility theory in various other ways, too, and, just like SC's preferences, some of these violations do not seem intuitively irrational. For instance, as we saw, there does not seem to be anything obviously wrong with the preferences displayed in the Allais paradox.

We may also want to challenge more directly the idea that any rational risk aversion must be explained by the concavity of the utility function. As we mentioned before, if we think of utility in the realist sense, as measuring the strength of our desire, it seems like we can be risk averse with regard to goods for which our utility is not diminishing. And two persons may agree on both the value of outcomes and on the relevant probabilities, and nevertheless evaluate a risky act differently - precisely because they have different attitudes to risk.[6] Alternatives to expected utility like REU theory hold out the promise of offering a treatment of risk aversion which can make better sense of the ordinary decision maker's attitudes to risk, while still being systematic. As I want to show now, Buchak's REU theory can apparently make sense of SC's preferences.

6.3 Risk-Weighted Expected Utility Theory

In expected utility theory, only two components go into the agent's choice rule, or the representation of the agent's preferences: probabilities and utilities. REU theory adds a third element, namely a risk function $r(p)$. It represents agents as maximizing their risk-weighted expected utility, or to act as if they were maximizing some risk-weighted expected utility function. According to Buchak, the REU of a gamble is meant to represent the instrumental value the agent assigns to a particular gamble. It tells us how good a means the agent thinks the gamble is for satisfying her various desires. As we have seen, Buchak, like us, takes the relevant desires to have features of outcomes only as their objects.

The three components of the REU representation just mentioned are to be interpreted as follows: The utility function is meant to capture the agent's desires over outcomes. As in expected utility theory, the probabilities capture her beliefs about how likely the various outcomes, or states that lead to those outcomes, are. And lastly, differences in the risk function represent differences in how agents structure the attainment of their goals. Two agents with the same utility and probability functions can thus disagree about the instrumental value of a gamble if they have different risk functions, and thus structure the attainment of their goals differently.

REU theory requires us to order the various different outcomes an act could lead to according to their utility. Let these be $x_1, x_2, \ldots x_n$, where the index indicates the outcome's rank in terms of utility, starting with the lowest. The REU of the act is then calculated as follows: To the utility of the worst outcome, we add the weighted utility difference between that outcome and the second worst, and then

[5] Note that if we think of utility as a real quantity measuring degrees of desire, then this implication in itself is implausible, quite apart from what it implies about the agent's level of risk aversion.

[6] See Buchak (2013), Chapter 1 for this line of critique.

the weighted utility difference between that outcome and the next better one, and so on.[7] The weights are a function $r(p)$ of the probability of receiving at least the utility of the higher of the two outcomes (see Buchak (2013), p. 53). Formally, the REU of a gamble g is give by

$$\begin{aligned} REU(g) =& u(x_1) + r(\sum_{i=2}^{n} p(x_i))(u(x_2) - u(x_1)) + r(\sum_{i=3}^{n} p(x_i))(u(x_3) - u(x_2)) \\ &+ \ldots + r(p(x_n))(u(x_n) - u(x_{n-1})) \end{aligned}$$

According to REU theory, decision-makers maximize their REU. Or, according to the constructivist stance we have favoured, agents behave as if they did, by having preferences that are representable as REU maximizing. Buchak provides a representation theorem that shows that agents whose preferences abide by a number of axioms, including a weaker version of separability[8] will be representable as REU maximizers. Most importantly for us, such agents can be risk averse even if we do not assign decreasing marginal utility in money to them, e.g. if we assign them linear utility in money.

Agents with linear utility in money turn out to be generally risk averse[9] when $r(p)$ is smaller than p, and convex. This means that, as the probability of receiving at least a particular outcome increases, the weight assigned to it becomes ever larger. An example of such a risk function Buchak uses throughout her book is that of $r(p) = p^2$. Given such a risk averse risk function, outcomes that are very good but unlikely will contribute less to the overall instrumental value of a gamble than they would for a less risk averse agent, or indeed for an expected utility maximizer. The risk function is meant to represent the agent's commitment to weighing the chance of receiving some outcomes more or less than others, depending on their rank.

Note that, unlike the equivalents in other rank-dependent utility theories, the risk function in REU theory is not meant to represent anything about the agent's subjective degree of belief, such as her taking an outcome to be more or less probable than its objective probability depending on its rank. Instead, Buchak takes it to be a more or less stable character trait that concerns how we act in the face of risk. At one point (pp. 55-56), she suggests that the risk function describes where the agent falls between the two virtues of prudence and venturesomeness.

REU theory is permissive about what attitudes to risk an agent may have while keeping the utility function representing her evaluation of outcomes, representing her desires, fixed. That is, it is permissive about how an agent may serve her desires in the context of uncertainty. She is free to have a wide range of different risk functions, or to be representable as such.

This, however, has the interesting consequence that REU theory cannot result in an instrumental requirement to act in accordance with one's preferences over prospects, or to maximize with regard to the risk function that represents one's preferences. This is because on REU theory's own terms, one could have had a different risk function, and been equally instrumentally rational. As long as one acts in accordance with some risk function that one could have had, even if one does not have it, it seems like one is not instrumentally criticizable. It is thus in fact misleading to claim that REU theory requires

[7] This makes REU theory a rank-dependent utility, similar to that of Quiggin (1982). In fact, the argument in this chapter also applies to Quiggin's theory.

[8] To be very brief, Buchak requires what she calls comonotonic trade-off consistency, while expected utility theory requires unrestricted trade-off consistency. Unrestricted trade-off consistency is a form of separability, and Buchak restricts it to apply only to prospects that order all states, and all events in the same way in terms of the agent's preferences over the outcomes or sub-prospects resulting from those states and events (see p. 98).

[9] Buchak defines general risk aversion as follows: For any two gambles g and h, where h is a mean-preserving spread of g, the agent prefers g (pp. 21-22).

agents to maximize their REU. We argued in Chapter 4 that there is often no one unique way to trade off one's desires even in the context of certainty, and that this means that *maximization* is no longer a requirement of instrumental rationality. REU theory's premise is that it is permissive about what attitudes an agent may have over prospects given the very same desires for outcomes. Hence, likewise, it cannot ground an instrumental requirement to act in accordance with the preferences over prospects one happens to have.

Given this is so, for REU theory to serve its explanatory or predictive function, it should interpret preferences in such a way that agents necessarily, or at least ordinarily act in accordance with them, that is, as dispositions to choose. This also seems apt given that Buchak interprets the risk function as a kind of commitment to treating risk in a certain way. REU theory then models agents as having choice dispositions that are representable as REU maximizing, where the risk function in that representation captures the agent's prudence, or venturesomeness, or commitment to a certain treatment of risk.

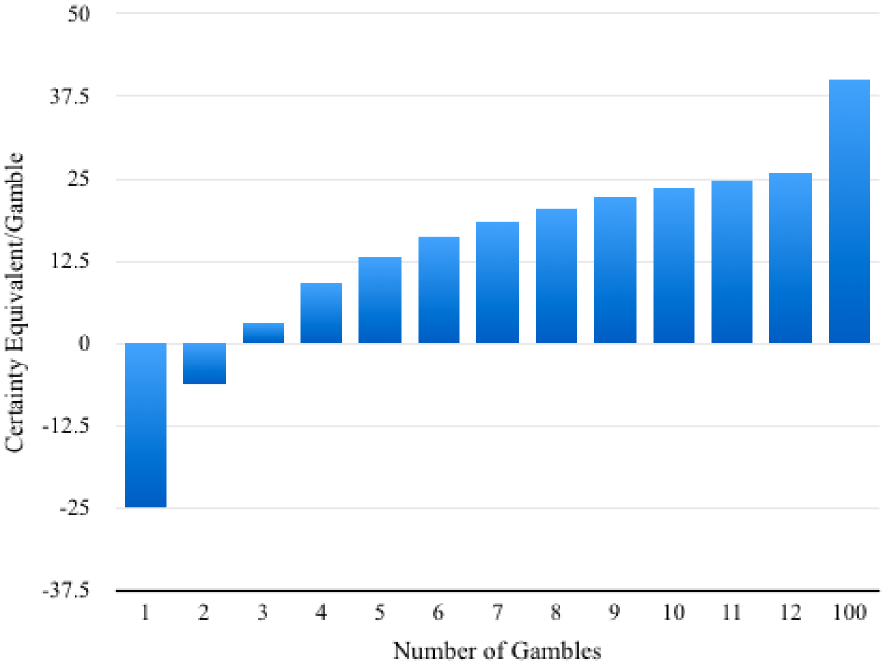
There are two ways we could now think of the point of REU theory. On the one hand, we could think that it is an expression of weaker requirements of instrumental rationality. In particular, we could view it as expressing a commitment to the weaker form of separability featured in Buchak's representation theorem as a principle of instrumental rationality. I have doubts, however, that this requirement will stand up to scrutiny as a principle of instrumental rationality. Any potential defence of even a weaker version of separability would seem to be subject to the same arguments we made in the case of separability in the last chapter — that is, unless that weaker version of separability is in fact state-wise dominance itself.

According to the other way of thinking about REU theory, the point of REU theory is that with the risk function, it allows us to represent the agent as having some stable disposition to choose in the face of risk. The REU representation aims to capture some real character trait in the face of risk, such as venturesomeness or prudence. And then the REU representation could serve some real explanatory or predictive function. The axioms of the representation theorem then merely specify the conditions under which such a representation is possible. Even if violating these conditions may not be instrumentally irrational, if most agents don't violate them, then REU theory could still offer a treatment of risk aversion that is predictive, explanatory, and casts it as instrumentally rational. I take it that this is the main motivation behind Buchak's theory.

Regarding the counter-examples to expected utility theory, the hope is then to not only show how it could be instrumentally rational to have the preferences displayed in them. The hope is to also explain these preferences in terms of some stable disposition to choose in the face of risk, or a commitment to 'structuring the attainment of one's goals' in a particular way. Buchak in fact applies the theory to various counter-examples to expected utility theory with that aim. The Allais problem we dealt with in the last chapter is one such example. Buchak shows that there are plausible combinations of risk, utility and probability functions that could result in the Allais preferences, given that the outcomes are well described by the different wealth levels the agent could achieve in the Allais lotteries (p.71).

Importantly for us, it also seems like REU theory can make sense of preferences like those of SC. Again assume that SC faces a one-off choice where the outcomes are adequately described by the different wealth levels the agent might achieve through the gamble he is considering taking. Suppose, for instance, that SC is an REU maximizer with linear utility in money, such that $u(m) = m$, where m is money, and risk function $r(p) = p^2$. The REU of accepting the gamble turns out to be -25.[10] Given that this is negative, he will reject the gamble.

But it turns out that the REU of accepting one hundred such gambles is 4004.56 given these utility and risk functions. Given that this agent's utility function is linear in money, this value can be interpreted as the certainty equivalent of the gamble: The agent is indifferent between receiving $4004.56 for certain, and the larger compound gamble. In fact, the agent would accept the compound gamble resulting from any series of three or more gambles. Figure 6.2 shows the certainty equivalent, divided by the number of individual gambles, of compounds resulting from series of gambles of various lengths.



.

Figure 6.2: Certainty Equivalent Per Gamble for Repetitions of SC's Gamble

Note that the certainty equivalent per gamble increases with the number of repetitions that are offered. In order for SC's preferences to turn out rational, this must in fact be so. After all, the certainty equivalent of an individual gamble must be negative for SC to reject it. But the certainty equivalent of the compound gamble must be positive for him to accept it, and so the certainty equivalent per individual gamble must also be positive. In the case of REU theory, it can even be shown that for any gamble, under the conditions we specified, as the number of repetitions increases to infinity, the REU certainty equivalent per gamble approaches the expected utility of the gamble (see Buchak (2013), pp. 217-18).

It helps to stop and consider how this is possible. For the decision-maker we considered, the individual gambles are independent in two senses: They are probabilistically independent, such that winning one of the gambles does not make winning the next one more likely. And they are independent in terms of

[10] Call SC's gamble s. We then have:

$$REU(s) = -100 + 0.5^2 \cdot 300 = -25$$

utility, such that winning one does not affect the increase in utility she might gain through the next. This is due to the linearity of the agent's utility function in money. As her wealth decreases or increases, she nevertheless receives the same amount of added utility from winning or losing the next gamble. Under these circumstances, the expected utility of a combination of gambles is just the sum of the expected utilities of the individual gambles.[11] And so for an expected utility maximizer for whom the different gambles are independent both in terms of utility and in terms of probability, each individual gamble would have the same value, whether it was part of a series of gambles or not.

This is not so for our REU maximizer. For her, there is a kind of interdependence or complementarity between the individual gambles that causes the instrumental value of the compound of the gambles to be different from the sum of the values of the individual gambles. Indeed, any theory which hopes to cast SC's preferences as rational must account for the interdependence of the instrumental values of the gambles he is offered. SC's gambles are probabilistically independent, and the failure of expected utility theory to account for SC's preferences suggests that interdependence in utility can't be the whole story: Decreasing marginal utility cannot explain SC's preferences, and it is not clear what other types of interdependence in utility could be involved here.

In the case of REU theory, it is the risk function which accounts for the complementarity in the instrumental value of the individual gambles. Due to the risk function, for an REU maximizer, the position of an outcome in the overall spread of outcomes, as well as the probability of other outcomes matters for how much one outcome contributes to the overall value of a gamble. And this is how compounding can lead to complementarities. It can affect the value contribution each of the outcomes of the individual gambles makes, by affecting the position of an outcome in the overall ranking of outcomes.

6.4 Risky Choice over Time

We have now seen that REU theory seems to be able to make sense of SC's preferences, while expected utility theory cannot. This seems to count in favour of REU theory. In the following, I want to show that the argument we just gave relies on ignoring an important observation, however. And that is that every individual risky choice we face in our lives is embedded in a long series of risky choices. This will also be so for any real-life case that resembles the choice SC faces. As we will see, given a commitment to resolute choice, and a plausible assumption about the specification of decision problems, this means that REU theory cannot in fact explain SC's choices.

We often face risky choices that resemble the individual gamble SC faced. We face similar monetary gambles when we decide how to invest our savings. But many non-monetary gambles also have a similar

[11] Take two gambles $g = (p_1, x_1; p_2, x_2)$ and $h = (p_3, x_3; p_4, x_4)$, where $p_1 + p_2 = 1$ and $p_3 + p_4 = 1$, the probabilities of the two gambles are independent, and the utility the agent gets out of one outcome is independent of the other outcomes she receives. In that case, we can show that the expected utility of the compound gamble gh is just the sum of the expected utilities of the individual gambles:

$$\begin{aligned}
EU(gh) &= p_1p_3u(x_1 + x_3) + p_1p_4u(x_1 + x_4) + p_2p_3u(x_2 + x_3) + p_2p_4u(x_2 + x_4)\\
&= p_1p_3(u(x_1) + u(x_3)) + p_1p_4(u(x_1) + u(x_4)) + p_2p_3(u(x_2) + u(x_3)) + p_2p_4(u(x_2) + u(x_4))\\
&= (p_1p_3 + p_1p_4)u(x_1) + (p_2p_3 + p_2p_4)u(x_2) + (p_1p_3 + p_2p_3)u(x_3) + (p_1p_4 + p_2p_4)u(x_4)\\
&= p_1u(x_1) + p_2u(x_2) + p_3u(x_3) + p_4u(x_4)\\
&= EU(g) + EU(h)
\end{aligned}$$

structure. Take again my choice of whether to cycle to work or take the subway. Suppose cycling always takes the same amount of time, and I can be certain that I will make my first appointment on time. The subway is less reliable: Sometimes, there is a train right as I step onto the platform. In that case I will even have time to get coffee before my appointment, which I would like very much. But sometimes, I have to wait for a long time, so that I am late. Or suppose that Steady Stan's Steaming Stews reliably produces bland food, and food at Fickle Freddy's Food Follies varies, with a 50% probability of either great or unpleasant food. My decision of where to go for lunch then has a similar structure, with a safe option and a riskier one which may have a better expected outcome.

Like in SC's case, it does not seem crazy to go for the safe option in these cases. But unlike SC's case, we rarely face a choice involving the corresponding compound gamble representing, for instance, eating at Fickle Freddy's Food Follies one hundred times. Seeing that decisions involving such compound gambles are rare, we might think that SC's case is of little relevance for ordinary decision-makers. But I think this is too hasty. I have to decide how to get to work, and what to have for lunch every day. I face these choices repeatedly. And that means that I will face the relevant compound gamble, e.g., eating at Fickle Freddy's Food Follies one hundred times, *over time*, even if I never face it in a one-off choice.

When we formulated SC's decision problem in the last section, we did not explicitly embed SC's choice in a long series of similar choices. We modelled his choice as a one-off decision, where the relevant outcomes are the wealth levels he can achieve through this one-off decision. Given this formulation of the decision problem, REU theory can, and expected utility theory cannot account for SC's preferences. But if SC really does face similar gambles repeatedly over time in any realistic scenario, we could also model the decision problem in a way that makes it explicit that each individual risky choice is embedded in a long series of risky choices SC faces in his life.

For now, I want to restrict the discussion to the monetary gambles we all face repeatedly. I will suppose that SC faces his gamble repeatedly over time. In the last chapters, we modelled cases of consecutive choice in dynamic choice problems. We can do the same here. Figure 6.3 illustrates the case where the agent only makes two consecutive choices in a row.

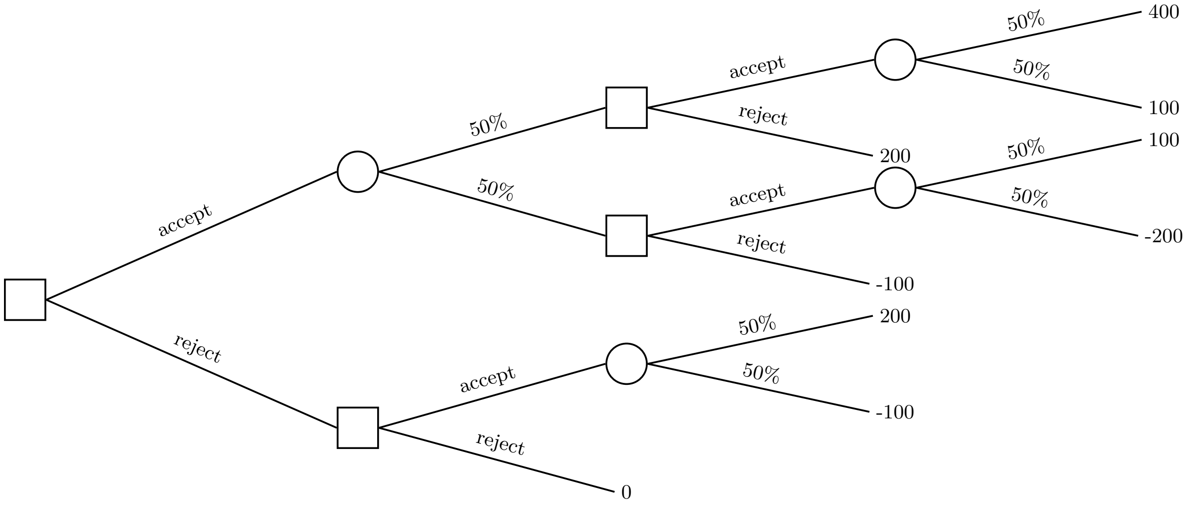


Figure 6.3: Dynamic Version of SC's Decision Problem, Two Consecutive Choices

As we saw in the last chapter, those defending alternatives to expected utility theory typically respond to the dynamic arguments we considered there by advocating some form of resolute choice.[12] Resolute agents decide within dynamic choice problems as they would were they to make a decision on their entire course of action in the dynamic choice problem at once. That decision, in turn, is guided by the agent's preferences over the prospects that are attainable in the dynamic choice problem as a whole.

As I now want to argue, if we insist that in any dynamic choice problems, the agent must be guided by her preferences over entire courses of action, as resolute agents are, REU theory can no longer account for SC's preferences when the decision problem is formulated in a way that makes it explicit that SC faces a long series of gambles in his life. And this is because, when the dynamic context is made explicit, the resolute agent's preferences over series of gambles come to matter. In SC's case, this means that the agent's preferences over the larger compound gamble that is faced over time come to matter. This will make him reject his gamble on the individual occasions he faces it.

Suppose SC faces his gamble one hundred times in a row, on one hundred consecutive days. How would a resolute agent decide in this case? She would consider her preferences over the entire courses of action open to her, and act in accordance with one of her most preferred courses of action. I assume that an agent like SC will have preferences over entire courses of action that reflect the preferences he has over the compound gamble we looked at previously.[13] In the case where SC's gamble is faced repeatedly, we thus get the result that he prefers the course of action of accepting the gamble in a series of choices, while preferring to reject a single gamble. The resolute agent would thus end up accepting the gamble every time.

For agents who violate the core requirements of expected utility theory, as we have seen, being resolute is a way of avoiding instrumental irrationality over time. That is, these agents avoid a sure loss of something they desire, such as money. Resolution is accordingly defended by various critics of expected utility theory, such as McClennen (1990), Machina (1989), and, to some extent, Buchak (2013).[14] We can in fact make a similar 'sure loss' argument in the context of SC's preferences.

As we have seen, the main alternative to resolute choice is sophisticated choice, which is usually associated with backward induction reasoning. We start at the end of the tree, and assume that the agent makes his decision based on his given preferences over the prospects associated with the final nodes. We then reason backwards from there. We eliminate those branches that the agent would not choose, and consider what the agent would decide at the previous choice nodes given that restriction, and given he is again guided by his preferences over the prospects available to him then, and so on.

In the tree in Figure 6.3, where only two choices are involved, at all the last choice nodes, the agent would choose to reject the gamble and go 'down', both given the REU function we have been using, and Samuelson's assumption that SC would reject the gamble at any wealth level he might be at during the series. But that means that at the first choice node, the agent will think of himself as just facing the one-off single gamble. Again, he will reject, and choose 'down'. If we iterate this for the case where the agent makes one hundred consecutive decisions, we get the same result. Unlike the resolute agent, the sophisticated agent ends up rejecting each gamble.

[12] As before, I here include Rabinowicz's unified choice, since it advocates the same kind of choice behaviour in all the cases we are considering.

[13] When we think of an agent who faces SC's original gamble repeatedly, this makes sense, because we are dealing with money. Money is usually not spent on the day it is earned, and thus will just add up over time to contribute to the agent's overall wealth. The timing of when he receives the money should not matter in a significant way, and thus the preferences over the course of action with regard to a series of choices involving SC's gamble should reflect preferences over the compound gamble over wealth levels the series results in.

[14] Buchak in fact keeps open the possibility that it may still be rational for REU maximizers to be sophisticated. However, we have argued that sophisticated agents are instrumentally irrational for violating principles S and T.

Now suppose that before the one hundred consecutive choices, SC is given the chance to bind himself to the one hundred gambles, but for a small fee. Being sophisticated, the agent would choose to pay this fee. After all, he knows that if he doesn't do so, he will reject each. His preferences over the prospects available at the beginning favour accepting all gambles, and if this preference is strict, he will be willing to pay at least a small fee to bind himself to accepting the gambles. In fact, if he is an REU maximizer with the risk and utility functions described earlier, he will be willing to pay \$4004.56. But then no matter how much he receives in the series of gambles, he will have done that much worse compared to the course of action whereby he does not bind himself, but still accepts all the gambles. This, we have argued, is irrational. An agent can avoid this sure loss by being resolute. The resolute agent will not pay the fee, but still accept all gambles.

It thus seems like the agent is well advised to be resolute. Moreover, given the permissiveness of REU theory regarding an agent's attitudes to risk, resolution does not appear to violate instrumental rationality — resolution merely seems to consist in making different individual choices in the context of the dynamic decision problem than one normally would. But as long as each of these individual choices are equally permissible given the agent's desires over outcomes, the agent can't be instrumentally criticized for that.

Allowing for resolution does mean, however, that we may have to restrict what preferences are permitted when constructing an REU representation of the agent's preferences. As we have seen, one way of interpreting resolution has the agent temporarily adjust her preferences within dynamic choice problems only. This is in fact our preferred notion of resolution, since it is the version we need if preferences are understood as choice dispositions. But these temporarily adjusted preferences are presumably not permitted when constructing an REU representation of the agent's preferences. This representation should capture what we called the agent's given preferences. If it does not, then the temporary adjustment of preferences in the dynamic choice problems would actually result in temporary changes in the risk function — and then the risk function would be ill-suited to capturing an agent's stable commitments or character in the face of risk.[15]

In the following, I want to argue that appealing to the strategy of resolution does not help in defending REU theory and similar alternatives to expected utility theory, given a common assumption about how agents ought to specify their decision problems. The next section will argue that for resolute agents who are REU maximizers, or have SC-type preferences, there is a disagreement between what an agent chooses when she takes different temporal perspectives resulting in different specifications of her (dynamic) choice problem. I will then argue that a common account of how agents ought to specify their decision problems implies that these agents should frame their decision problem in the widest way possible. If they do so, however, they end up acting approximately like expected utility maximizers. This also makes it the case that REU theory cannot account for the preferences exhibited in the common counter-examples to expected utility theory after all. That is, they cannot account for the examples that motivated them anymore.

[15] Briggs (2016) in fact suggests that REU maximizers should respond to the dynamic choice arguments of the last chapter in this way. But for the reason just stated, this seems to go against the spirit of REU theory.

6.5 Framing Decision Problems

Consider again the repeated choices I described SC as facing in the last section. I described him as facing one hundred individual SC gambles on one hundred consecutive days. We then looked at a dynamic choice problem which only included these one hundred choices. But why should we focus only on the choices that SC faces within those one hundred days? And why should we only focus on the choices he faces with regard to the SC-type monetary gambles? We said that we face many sorts of risky decisions many times in our lives. Each future such decision could potentially be modelled explicitly when thinking about what to do for the next small stakes risky choice.

Given that my life consists of repeatedly taking decisions with uncertain outcomes, I also face the choice of how to formulate my decision problems: How many and which of my future decisions should I take into account in the formulation of my decision problem? Or, to use Buchak's terminology, how should I 'carve out' my decision problem? We may also think of this as a question of temporal perspective. SC might take the perspective of an entire year, or he could only look at the next week, or the next day. Beyond his individual monetary gambles, he might consider all the risky choices he is making in a day, or only those risky choices of a particular type, like those concerning food. These different perspectives result in different ways of formulating his choice problem. And, it follows from the foregoing that given his preferences and the fact that he is resolute, his choice of perspective will matter for what he concludes he should do.

Any resolute agent with SC's preferences will evaluate an individual gamble differently depending on whether she conceives of herself as choosing it in a one-off choice or as part of a series of one hundred similar choices. And presumably, for such an agent, embedding a gamble in an even larger sequence of gambles will further change her evaluation of the single gamble.[16]

So how should a resolute agent with SC's preferences carve out his decision problems? This is not a choice that a formal decision theory like REU theory can help the agent with, since it is a decision the agent has to make prior to applying REU theory. Perhaps we could say that there is no choice of framing for agents to be made. They just approach decision problems from a particular perspective and proceed from there. As long as they make their decisions rationally given this perspective, their choice was rational. But I think this is not convincing. At least when contemplating important decisions, we often look at those decisions from different perspectives. In our deliberations, we may wonder how a decision fits into our plans for the week or the year or the month. It seems that we can deliberately choose which recommendation to go with when making the final decision.

A common view on how agents ought to specify their decision problems is that they should include in their specification of their decision problems everything that is *relevant* to their decision. Moreover, every detail that would change the agent's decision if it was included is usually deemed relevant. In the following, I will adopt this notion of relevance. Note that, as a matter of fact, SC in our example will face further decisions in the future. And given his preferences, whether the gamble is faced again in the future makes a difference to his evaluations, and thereby seems relevant. If widening one's perspective

[16] In fact, there is empirical evidence that which of SC's preferences, the large stakes or the small stakes one, comes to play is indeed a matter of the temporal perspective agents choose to take. Benartzi and Thaler (1999) conducted surveys to investigate investment behaviour in retirement savings. What they found is that many who reject a portfolio when they are shown its annual returns will accept it when they are shown the resulting distribution over a long period of time. And note that, in the case where these agents are shown annual returns, it is not the case that they actually only face a one-off investment decision. Presumably all agents are interested in investing money over longer periods of time. But the different representations of the returns could plausibly be taken to trigger different temporal perspectives. And so what seems to make the difference is what temporal perspective these agents take when carving out their decision problems.

means taking account of something which is relevant, then it seems like we should do that. In fact, Buchak (2013) herself claims that a decision problem should include everything that the agent deems relevant to her decision (p.37). And so the most long-term perspective seems to be the appropriate one to take for our agent.

Following Savage (1954), a decision problem which specifies everything that could be relevant to the agent's decision is sometimes called a 'grand-world' decision problem. In the context of Savage's decision theory, Joyce (1999) introduces grand-world decision problems as non-dynamic decision problems which individuate the possible states of the world, the actions available to an agent, and the possible outcomes of those actions in such a way that no further refinement would change the agent's attitudes. We can also translate the notion of a grand-world decision problem into the context of dynamic decision problems. Here we could think of a grand-world decision problem as one that explicitly models all the agent's future decisions as part of the dynamic decision problem.

Due to the intractability of grand-world decision problems, in decision theory, we only deal with coarsenings of the grand-world decision problem, or 'small-world' decision problems. Joyce (1999) argues that decisions based on small-world decision problems are only rational to the extent that we are justified in believing that we would come to the same conclusion were we to consider the grand-world decision problem (p.74). The grand-world decision problem is the problem that rational agents should ultimately be trying to solve, but as limited beings, we use small-world decision problems that we hope are a good enough model of the grand-world problem.

In the case where SC is resolute and finds himself in a context where he faces his gamble repeatedly, including those future decisions in his decision problem will change what he thinks he should do. And thus he cannot be justified in believing that the solution to a small-world decision problem that only specifies a one-off choice will lead him to choose what he would have chosen in the grand-world decision problem. If SC is resolute, and aims to include everything that is relevant to his decision in the decision problem, then he should not consider himself as facing only a one-off choice. His preference for rejecting an individual gamble should therefore never come to bear. And then, if we think that SC actually finds himself in a context where he faces the gamble he is offered repeatedly, it would not be permitted for him to reject the gamble. Given the assumption that he actually faces his gamble repeatedly, we can suddenly no longer account for SC's choice behaviour.

We might think that all of this points to a very specific problem with SC's case only. However, as I want to argue below, I think that we often find ourselves in structurally similar situations, which raise the same problem.[17] Moreover, the case of SC merely serves as an example for a general problem for REU theory. This is what I will turn to in the next section.

[17] Interestingly, there also seem to be examples where the divergence between the evaluation of an individual risky choice and the evaluation of a series of such choices goes the other way, that is, we display more risk averse preferences on the larger scale. This is so when we repeatedly face a small chance of something very bad happening. Suppose that every day there is a 0.01% chance of me being hit by a car when I jaywalk on the way from the subway to work. I might reasonably think that that chance is too low for me to justify walking an additional 200 metres. But if I jaywalk every day, this amounts to a 3.6% chance of me being hit by a car this year. And I might very well think that this probability justifies me walking an additional 200 metres every day. In these cases, we will have difficulties salvaging the rationality of an agent who takes risks with a very small probability of something very bad happening. Tenenbaum and Raffman (2012) discuss a structurally similar example that involves smoking and the risk of developing cancer.

6.6 Risk-weighted Expected Utility Theory and Framing

SC has preferences which make future repetitions of his gamble relevant to a present choice he faces, given that he is resolute. As we have seen in Section 6.3, REU theory can apparently account for his preferences because it involves a risk function which introduces complementarities between different risky choices the agent faces, even when these are independent in terms of probability and utility. But this feature of REU theory means that the observations of the last section lead to a very general problem for REU theory.

Apart from the special case where $r = p$ and REU theory reduces to expected utility theory, the risk function creates complementarities between any two risky choices an agent faces in her life. This is because the risk function is applied after utilities have already been assigned. In REU theory, just as in expected utility theory, utilities are assigned to outcomes involving different kinds of goods, occurring at different times and places. Having such a single measure allows us to express the way in which the agent trades off different kinds of goods, or the way in which she evaluates gambles that involve different kinds of goods. The risk function is applied to this single utility measure. And thus it creates complementarities between any two gambles the agent faces. We all face many risky choices of different kinds in our lives, and so this feature of REU theory has far-reaching consequences.

We can illustrate this with the transport and restaurant choice scenarios introduced above. Like any risky choices, according to REU theory, these can also be described as choices over utility gambles. Suppose that the probabilities involved in these gambles are also independent for any two occasions where I face the choice. Moreover, the agent's preferences over the outcomes are such that they need to be modelled with utilities that are independent: Whether I took the subway one day does not affect how much added utility I get from taking it the next day, and whether I went to Steady Stan's one day does not affect how much added utility I get from going there the next day. Further suppose that the utilities in the transport and restaurant choice are the same as those in SC's original choice.

Now assume that I am resolute in dynamic decision problems, and can be assigned the utility and risk functions we used above. In that case, it follows from what we have said that if I think of my choices of means of transport two at a time, I will come to a different conclusion from when I think of them three at a time. In the first case I will cycle both times; in the second case I will take the subway each time. Or suppose I decide to consider my lunch decision together with my decision of what means of transport to take on two consecutive days. This will make me change my choice from cycling to taking the subway, along with eating at the riskier restaurant - even though the utilities of the possible outcomes of all these decisions are independent from one another.

The fact that REU theory creates complementarities between all the risky choices an agent faces, together with the assumption that the agent is resolute thus leads to a serious problem for REU theory. It ensures that an agent can never be justified in believing that the solution to a decision problem that falls short of taking account of all the future decisions she expects to face is going to approximate the solution to the grand-world decision problem.[18] For the REU maximizer, different temporal perspectives and the corresponding formulations of choice problems may result in different recommendations regarding one

[18] Buchak herself (pp. 228-229) seems to think that REU maximizers should package together decisions that concern the same kind of value, and may separate decisions that concern values that are unrelated to each other into different decision problems. However, this strategy would lead to an under-specification of decision problems according to Joyce's rule, which Buchak also seems to be committed to. This is because the risk function creates a complementarity between risky choices even if they concern different kinds of values. REU theory itself makes risky gambles involving different types of values complementary to each other.

and the same action, even if that action will influence neither the probability nor the utility of any other outcomes she may end up with in the future. More short-term perspectives will not generally agree with the most long-term perspective.

Note that expected utility theory does not have this problem. For the expected utility maximizer, when she faces a series of gambles which are independent in terms of utility and probability, how she carves up the decision problem is not going to matter. In that case, the expected utility of the compound gamble is the same as the sum of the expected utilities. When she constructs a small world decision problem in which she faces a one-off choice of whether to accept an individual gamble, this will give her the same conclusion as the one she would draw were she to consider the series of gambles as a whole.

We have seen that a resolute REU maximizer can apparently never be justified in believing that a choice based on a decision problem which considers only a subset of the gambles she will face in her life is going to cohere with the decision she would make were she to consider the grand-world decision problem. If we think that agents are required to formulate their decision problems to ensure that there is such a coherence, then it seems like she is forced to think about her grand-world decision problem every time she makes a choice. This would make REU theory an even more demanding theory to apply than expected utility theory.

What makes things worse is that, if I am a resolute REU maximizer and I want to make the decision that I would come to were I to consider the grand-world decision problem, it seems my safest bet is to behave as if I were an expected utility maximizer. We have seen above that for a risk averse REU maximizer, as the number of repetitions of a gamble goes to infinity, the average certainty equivalent of each gamble tends to its expected utility. And so a resolute agent considering a large number of repetitions of the same gamble will choose just like an expected utility maximizer. Now of course the actual grand-world problem an agent faces will be more complicated than a large number of repetitions of the same gamble. But we can speculate that the results should be similar for large numbers of different gambles.[19] And so resolute REU maximizers will in fact end up behaving approximately like expected utility maximizers.

This conclusion thus undermines the central motivation for REU theory, namely that it can account for various counter-examples to expected utility theory. Assuming the agent is resolute, and considers herself as facing the grand-world decision problem, REU theory cannot in fact cast choices like those of SC as rational. When SC says that he would reject the single gamble, he apparently fails to integrate his decision concerning that individual gamble into the grand-world decision problem — or otherwise he would accept it.[20] As Benartzi and Thaler (1999) point out, he was probably already playing less favourable gambles every day by holding some of his retirement savings in stocks.

In fact, all of the toy examples that are used to motivate and illustrate REU theory, and not only SC's case, involve agents who must have left out relevant detail from the specification of their decision problems, given they are resolute. These examples usually involve one-off choices involving outcomes that

[19] For one, Buchak's repetition theorem will also apply, for instance, to the compound gamble resulting from the many choices one makes in a typical day. Another supporting consideration is that, as long as different gambles are reasonably independent, an agent's exposure to risk throughout her life is diversified, and the variance of the overall risk 'portfolio' is lower than the weighted average of the individual gambles' variances. The lifetime gamble is in that sense less risky than the constituent gambles. The risk function makes the agent sensitive to that lower variance.

[20] In a similar vein, Benartzi and Thaler (1995, 1999) refer to risk aversion for small stakes gambles as "myopic loss aversion". What they suggest explains these preferences is that when agents are shown the rates of return for an individual gamble, they are loss averse in the way Kahnemann and Tversky (1984) describe it. And agents fail to integrate the small loss into their current wealth level, which Benartzi and Thaler interpret as an instance of "narrow framing", as in Kahnemann and Lovallo (1993).

are described merely as the immediate goods one receives as the consequence of having made that choice. For instance, in the Allais problem described in the last chapter, the agent receives a sum of money with a particular probability as a result of her choices. Buchak shows how REU theory apparently makes sense of the choices typically made in the Allais problem, given that the agent's risk function is convex (p.71). She describes the outcomes simply as those sums of money the agent stands to gain through the gambles she faces. This is not integrated with all of the future uncertain monetary prospects the agent faces. But if the agent is a resolute REU maximizer who faces future risky choices, and his decision problem should include everything relevant, then it should be so integrated.

Parallel claims apply to the other examples used to motivate REU theory. Given that these examples are usually presented as simple small-world problems, we can no longer be sure that on a specification of the decision problem that includes everything that is relevant to a resolute agent, REU theory can make sense of them. In fact, if these problems involve small stakes gambles and we think that the agent faces many such gambles in her life, then, as we have said before, a resolute agent should behave roughly like an expected utility maximizer.

This conclusion also raises the question of whether, quite apart from its treatment of these motivating examples, REU theory offers a true alternative to expected utility theory. After all, given our assumptions, REU theory and expected utility theory license the same choices. Indeed, out of the two, expected utility theory is the simpler theory, making it more attractive. One might think that there is still a difference since expected utility maximizers and resolute REU maximizers behave in the way they do for different reasons. For an REU maximizer, it is only because a risky action like accepting SC's gamble is part of a long series of risky actions that she chooses to do it. If she faced the choice in isolation, she wouldn't. But since we generally all face more than one risky choice in our lives, this hypothetical difference is a very remote possibility.

We may then point out that the REU maximizer's given preferences still violate expected utility theory, even if she does not act on them. This observation, however, does not help REU theory. On our preferred way of interpreting resolute choice, the agent temporarily adjusts her preferences within a dynamic choice problem only. Now note that if the agent considers herself as facing one grand-world dynamic decision problem in her life, that may in fact amount to adjusting *all* her preferences.

We said that the REU representation of an agent will be based on her given preferences. But it is no longer clear what the relevance of given preferences even is if they may never manifest themselves in choice. Given preferences presumably capture how the agent would choose outside of the context of a dynamic choice problem at hand. But in the case we are considering here, the one grand-world decision problem is all there is for the agent. And in that problem, there is no difference in the actual preferences of REU maximizers and expected utility maximizers. In the light of the last chapter, it is especially important to note that a resolute REU maximizer who considers the grand-world decision problem ends up adopting separable preferences after all.

Most importantly, this means that any REU representation we may still assign to the agent's given preferences can no longer capture what REU theory set out to capture. The distinguishing feature of REU theory was supposed to be that it introduces a risk function that represents the agent's commitment to treating risks in a certain way, or her venturesome or prudent character traits. A resolute REU maximizer who acts like an expected utility maximizer in the grand-world decision problem displays no such character traits or commitments in her choices, since the given preferences that may be representable as REU maximizing never manifest themselves in choice. Under those circumstances, we do not seem licensed to interpret any risk function we may assign to the agent as representing a disposition, commitment or character trait.

These conclusion can be avoided, however, if we give up either the assumption of resolution, or the assumption that everything that is relevant to the agent's decision should be included in her decision problem. The next section considers the implications of relaxing these assumptions. It will turn out that relaxing the assumption of resolution, at least, is not instrumentally irrational. However, relaxing these assumptions also means that an REU maximizer's choices will be very sensitive to her choice of temporal frame and choice behaviour in dynamic decision problems. This sensitivity again stands in the way of REU theory identifying any stable choice disposition, commitment or character trait in the face of risk.

6.7 Permissiveness About Dynamic Choice

We have seen that REU maximizers will choose roughly like expected utility maximizers given two conditions: First, the agent is resolute, and chooses in dynamic decision problems as she would, were she to make one choice outright. Second, the agent includes everything in her dynamic decision problem that she deems relevant to her choice, where any detail that would change the agent's choice were it included is relevant. Given these two assumptions, REU maximizers should consider themselves as facing a grand-world decision problem including all of the choices they will face in their life. In that problem, they will behave roughly like expected utility maximizers.

The same will in fact hold for any alternative to expected utility theory that tries to make sense of SC's preferences. This is due to the fact that for anyone with SC's preferences, there exists a divergence between the evaluation of a gamble in isolation and the evaluation of a series of gambles. Seeing that SC-type violations of expected utility theory are common, any alternative to expected utility theory that tries to make better sense of our ordinary attitudes to choice under uncertainty should try to account for them.

One might think that this means that expected utility theory and the requirement of separability are on more solid ground than the last chapter has made it seem. While agents who violate expected utility theory can avoid instrumental irrationality by being resolute, resoluteness also makes them act roughly like expected utility maximizers if they consider their grand world decision problem. But does this mean that instrumental rationality requires agents to be expected utility maximizers after all? It only does so if instrumental rationality requires those who violate expected utility theory to be resolute, and to in fact consider their grand-world decision problem.

Alternatives to expected utility theory like REU theory can avoid the conclusion we reached in the last section by either relaxing the assumption that agents should be resolute, or by relaxing the assumption that anything relevant ought to be included in the choice problem, or indeed by relaxing both conditions. However, both conditions are commonly accepted by those advocating alternatives to expected utility theory. And so either relaxation would be revisionary. Indeed, as I want to argue here, if we give up these assumptions, the recommendations of REU theory become very sensitive to the agent's choice of temporal framing and her choice behaviour in dynamic choice problems. This sensitivity is an unwelcome consequence for REU theory.

If we wanted to relax the assumption that everything relevant ought to be included in the agent's dynamic choice problem, perhaps we could say that rationality is silent on the question of what perspective the agent ought to take. We may even hold that instrumental rationality is perspective-dependent in the sense that we can only ever declare an action rational with reference to whatever temporal perspective the agent took, while the choice of perspective itself is arational.[21]

One might support such permissiveness about temporal perspective by pointing out that if the inclusion of some factor in my decision problem changes my choice, this change need not mean that my new choice better serves my desires. It may just mean that I take another of two permissible choices. And then it seems like I wasn't rationally required to take that additional factor into account. Note that we appealed to similar reasoning when arguing before that sensitivity to framing need not be irrational. In fact, if REU theory wants to embrace being permissive about temporal perspective, then it needs to allow for sensitivity to 'temporal framing'. But Buchak herself considers sensitivity to framing in more traditional contexts irrational (pp.37-38).

It may not be true that every shift of perspective is rationally innocuous in terms of desire-based instrumental rationality. Still, for the sake of argument, suppose the agent is free to take whatever perspective she likes, even if this means leaving out 'relevant information'. Further suppose that the agent remains resolute. What would that mean for REU theory, or other alternatives that try to account for SC's preferences? It seems to imply that, for agents like SC, almost any course of action could be endorsed by the theory with reference to the right perspective. In our example, somebody with an REU function that can account for SC's choices in a one-off decision may accept all gambles or reject them all in a series of choices. The agent may even group different occasions in dynamic decision problems of different length, such that she accepts some but not others. Similar claims apply for any more complex series of gambles an REU maximizer may face.

Since REU theory started out as a theory that is permissive about choice under uncertainty, we may think that this is just a straightforward symptom of the theory. In fact, we have already argued that REU theory can't result in an instrumental requirement to act in accordance with the preferences over prospects one happens to have, given that other preferences would have been equally rational. In the light of that, the added benefit of REU theory was supposed to be that the risk function can capture an agent's stable disposition to take risks, or her 'venturesomeness'. What we have found now, however, is that the theory allows for a wide range of different choice behaviours even given some fixed risk function. An agent with an extremely risk averse risk function may still accept an SC-type gamble, namely if she is resolute in a grand-world decision problem. And an agent who has a much more 'venturesome' risk function may still reject the gamble, by thinking of each choice separately.

If this is so, it seems like the risk function no longer in fact captures any kind of stable disposition, or character trait. At most, it captures a disposition of prudence or venturesomeness relative to some temporal perspective. But, firstly, unless the agent also has a stable disposition to choose particular temporal perspectives, this will not result in any stable choice behaviour in the face of uncertainty independently of temporal perspective. And secondly, this still assumes that the agent is resolute. Given a temporal perspective and the assumption of resoluteness, an agent like SC must act in accordance with his preferences over the full prospects that may be brought about in the series of choices. Thus, given a temporal perspective, a resolute REU maximizer's choice behaviour is fixed, and we could think of the risk function as describing a perspective-relative disposition. If we give up the assumption of resoluteness,

[21] McClennen (1990) seems to suggest something like this when he writes the following in the context of the debate on how the outcomes of a decision problem should be specified: "If the world in fact opens to endless possibilities, still evaluation of risks and uncertainties requires some sort of closure [...] Wherever the agent sets his horizons, it is here that he will have to mark outcomes as terminal outcomes - as having values that may be realized by deliberate choice, but nevertheless as black boxes whose contents, being undescribed, are evaluatively irrelevant." (p. 249)

we can no longer even say that.

As we saw in the last chapter, resolution is advocated by many critics of expected utility theory because resolution helps an agent whose preferences violate expected utility theory avoid the instrumental irrationality of 'sure loss', or ending up unnecessarily frustrating one of one's desires no matter what happens. Could relaxing resolution then still be compatible with the instrumental rationality of an REU maximizer, or any other agent who violates expected utility theory?

In fact those who violate expected utility theory need not be resolute in order to avoid sure loss. We can already see this in the example of Rory presented in Section 5.4. Remember that Rory has Allais preferences and faces a dynamic Allais problem. Resolute Rory chooses not to bind himself, and goes 'up' at t_2 in correspondence with the preferences he has over the prospects initially available to him. But he would equally well avoid sure loss by choosing not to bind himself, and then going 'down' at t_2. He then ends up with a sure million, and thus not definitely worse off than if he had taken the risky gamble, with or without costs of pre-commitment.

Rory could achieve to go 'down' at t_2 by acting against his preference over the prospects available to him at t_0 — since his preferences then favour binding himself for a cost, given he correctly predicts he will go 'down' later. Or he could do it by temporarily adjusting that initial preference so that it favours going 'down' later. Either way, he will avoid sure loss, and violating principle S (see Section 5.7). However, he will not have chosen in accordance with his preferences over the initial prospects, as required by resolution.

We can in fact say the same about the sure loss scenario SC may face if he is offered a costly choice to bind himself. Being resolute is one way of avoiding the sure loss of paying (potentially a lot of) money to bind himself to accepting SC-type gambles in the future. But he also avoids sure loss by not binding himself and rejecting the gambles on each occasion, or on some occasions. As long as there is some chance that the agent would do worse by accepting more gambles, then this alternative response also avoids the instrumental irrationality of making a sure loss and violating principle S.

What we can say is that those whose preferences violate expected utility theory must in some respects fail to be sophisticated with regard to their given preferences in order to be instrumentally rational. Those who advocate alternatives to expected utility theory thus must allow for agents to sometimes choose against their preferences over the prospects still available, or to adjust their preferences temporarily within dynamic choice problems. In particular, REU theory must allow for agents to act against their given preferences at least some of the time if it wants to be compatible with the agent's instrumental rationality. However, there are many ways in which an agent might do so to avoid violating principles S or T.

If principles S or T are the only requirements that restrict the agent's choices under uncertainty, then it seems like we can indeed be very permissive about how the agent behaves in repeated SC-type cases, or in repeated choice under uncertainty more generally. One caveat could be the following: We have mentioned that taking extreme risks may not in fact be instrumentally rational. As we said, foregoing 99 muffins for a 50/50 chance of 100 does not seem like a good means towards the fulfilment of one's desire for muffins. Likewise, we might think that it may be instrumentally irrational to forego very safe bets, like foregoing one muffin for the same bet.

In SC's case, it might be that rejecting one hundred of SC's gambles, whether in a single choice or over time, crosses such a threshold to being instrumentally irrational. As we said, it is almost certain that SC won't make a loss over time, and likely that he will gain significantly from accepting each gamble.[22] If we think that this is enough to make him instrumentally irrational, however, the agent again need not accept every single gamble to avoid said irrationality. He just needs to accept enough to dispel worries about foregoing an almost certain gain.[23] And so even given this limitation, we could still be rather permissive about how agents act in repeated risky choice.

What would this permissiveness mean for REU theory? As with permissiveness about temporal perspective, it would mean that the same combination of risk, utility, and probability function could manifest itself in a variety of different choice behaviours. In this case, it could lead to a wide range of different choice behaviours even given a fixed temporal perspective. For example, any agent who has an REU function such that she would choose like SC in one-off decision problems may accept or reject all gambles she is offered over time, or indeed choose to accept only some, even if she considers herself as facing her grand-world decision problem.

What we have found, then, is that relaxing the assumption that agents are resolute, and relaxing the condition that she needs to consider herself as facing her grand-world decision problem makes the predictions of REU theory extremely sensitive to both the agent's choice of temporal frame, and the way she chooses to behave in dynamic decision problems. But this is a kind of fragility that REU theory wanted to avoid. As we said, Buchak takes sensitivity to framing, for instance, to be irrational.

Moreover, this sensitivity calls into question REU theory's claim to representing an agent's stable dispositions in the face of risk via the risk function. As we have seen, even given some fixed risk, utility and probability function, what an agent will do according to REU theory is still very sensitive to both the agent's choice of temporal frame, and her choice behaviour within dynamic choice problems.

The sensitivity we just described also in fact severely complicates the project of constructing an REU representation of the agent's preferences in the first place. We said before that if we take the risk, utility and probability functions to be mere constructs that represent the agent's preferences, then these should be constructs representing the agent's *given* preferences — those she has before potentially adjusting her preferences temporarily in a dynamic context. But in order to avoid violating principles S or T, agent's may be acting against their given preferences much of the time. The project of still coming up with an REU representation for the agent depends on being able to identify her given preferences. But that could be very hard in this context.

Given all of this, I think it is in fact no longer clear what the added benefit is of representing instrumentally rational agents as REU maximizers, rather than simply as agents who have much leeway in how to choose under uncertainty — as long as they abide by principles S and T, and do not run extreme risks over time, or forgo extremely good chances of benefit. If REU theory wants to be compatible with the agent's instrumental rationality over time, then it must allow her to not act on her given preferences, which are the basis of the REU representation, at least some of the time. If REU theory also wants to be able to account for choice behaviour like SC's, or the choice behaviour in other counter-examples to expected utility theory, then it can't require agents to be resolute as well as consider themselves as facing their grand-world decision problem. But if REU theory is permissive about the framing of decision problems and about when and how agents act against their given preferences, then it turns out to be sensitive to the agent's choice of temporal frame, and dynamic decision-making.

[22] Bovens (1999) argues that for this reason, avoiding small stakes risk aversion has instrumental value. Avoiding small stakes risk aversion enables an agent like SC who faces many gambles over time to reap almost certain benefits over time. Bovens argues that hope can play such an enabling role, and that this contributes to the instrumental value of hope.

[23] In fact, Tenenbaum and Raffman (2012) argue that it is a problem in expected utility theory that it can't explain why it may be instrumentally rational to adopt a policy of taking certain risks, such as smoking, on a limited number of occasions. On the permissive view we are describing here, agents are free to adopt such policies.

This, however, severely weakens REU theory's claim to describing a stable disposition, commitment, or character trait concerning how the agent structures the attainment of her goals in the face of risk. If the same REU representation is compatible with many different choice behaviours, it is not clear that the risk function can play an important explanatory or predictive role anymore.

6.8 Conclusions

The last chapter showed that separability, the core principle of expected utility theory, can only be defended as a requirement of instrumental rationality to agents who desire to have stable choice dispositions over time and across different choice situations. Agents who do not have such a desire seem to be free to have non-separable preferences. This chapter has considered whether an alternative formal decision theory that relaxes separability may be able to do better. I have focused, in particular, on Buchak's REU theory.

Appeal to a resolute choice strategy is a common move when defending alternatives to expected utility, such as REU theory, against the kind of dynamic choice argument we encountered in the last chapter. Here we have seen that a commitment to resoluteness, together with a common assumption about how agents should frame their decision problems implies that REU maximizers in fact behave roughly like expected utility maximizers. The same should hold for any other theory that tries to account for preferences like those of SC. Given the commitment to resoluteness, these theories thus do not offer a real alternative to expected utility theory. In particular, they cannot account for the counter-examples to expected utility theory that motivated them after all.

To avoid this conclusion, we need to either relax the assumption that agents need to take everything that is relevant into account when formulating their choice problems, or relax the assumption that agents need to be resolute, or relax both. Indeed, we have argued that at least the latter is permitted by instrumental rationality. Resoluteness is often treated like the only way for a non-expected utility maximizing agent to avoid the instrumental irrationality of sure loss in dynamic choice problems like those considered in the last chapter. However, as we have seen, it is in fact not the only such way.

Being permissive about the choice of temporal perspective, or about the way agents may choose in dynamic choice problems can help us account again for the choices displayed in standard counter-examples to expected utility theory. But this permissiveness also makes the predictions of REU theory very sensitive to an agent's choice of temporal frame, and her choice about how to treat dynamic decision problems. This, we have argued, undermines its claim to capturing a stable choice disposition in the face of risk.

In a sense, this should not surprise us. We saw in the last chapter that if an agent did aim to treat the same prospects in the same way at different times, and in different choice situations, then she should in fact have separable preferences. This is the only way of avoiding the instrumental irrationality of violating principles S or T, while at the same time not deviating from one's usual behaviour in dynamic choice contexts. Any alternative to expected utility theory must allow for such deviations in order to avoid violations of S or T. But given that we could potentially model the agent's entire life as one long dynamic choice problem with many opportunities for making a sure loss, it may be hard to distinguish the agent's 'normal' choice behaviour from the deviations, or to identify any stable disposition of how the agent tends to treat choice in the face of risk.

What can we say about REU theory, then? If its point was to merely articulate that instrumental rationality is permissive when it comes to choice under uncertainty, then it is right about that, and it does that well.[24] If, however, its point was to do more than that, and to capture risk aversion as a stable character trait, this chapter has shown that it fails.

[24] However, as we said above, there is a worry that it is still too restrictive. Any attempt to justify REU theory's weaker version of separability may fall prey to the same problems that the attempted justification of separability in the last chapter was subject to.

Chapter 7

Concluding Remarks

This dissertation started out by noting that most decision theorists are Humeans about decision theory. They believe that the requirements of decision theory are requirements of instrumental rationality, and of instrumental rationality alone. Most of the requirements that I have discussed here have considerable intuitive appeal as requirements of instrumental rationality. Of course, it seems, we should go with the option we most prefer. Likewise, money pump arguments seem to make it very plausible that agents should not have cyclical preferences.

One question that we should ask ourselves when faced with this intuitive appeal is what we implicitly take to be the standard of instrumental rationality. Which of the agent's conative attitudes are the standard by which we evaluate the instrumental adequacy of her actions? What I have tried to do in this dissertation is draw attention to that question. Once we consider it seriously, I argue, we find that the plausibility of different requirements of orthodox decision theory seems to rely on different answers to the question. The joint plausibility of the different requirements thus seems to rely on equivocation about the standard of instrumental rationality.

The different chapters of this dissertation have tried to show how this equivocation calls into question various instrumentalist arguments in the decision theory and dynamic choice literature, and ultimately makes the instrumentalist foundation of orthodox decision theory shakier than it may seem. The core problem seems to be this: Requirements that capture that one's actions should be guided by one's preferences over the objects of choice are most plausible if we take those very preferences to be the standard of instrumental rationality. On the other hand, instrumentalist arguments in favour of requirements on the *structure* of an agent's preference relation, such as acyclicity or separability, rely on the observation that agents who violate them may end up making a *sure loss.* The irrationality of this sure loss, however, is best cashed out by appealing to more basic desires over features of outcomes as the standard of instrumental rationality.

Chapter 2 discussed instrumentalist arguments in favour of resisting temptation, which are in fact arguments *against* one core requirement of orthodox decision theory, namely *maximization.* Most of this debate has taken the agent's preferences over outcomes to be the standard of instrumental rationality. It is for that reason that the instrumental rationality of resisting temptation is called into question in the first place: Resisting temptation requires us to act counter-preferentially, which seems instrumentally irrational according to that standard. This chapter shows that sticking with preferences as the standard of instrumental rationality, the instrumentalist arguments in favour of resisting temptation ultimately fail. They fail because they cannot establish that a strategy of resisting temptation is really in the agent's interests. In fact, the intuition that resisting temptation is rational really seems to be based on an understanding of the standard of instrumental rationality as something other than preference, something that preferences can be mistaken about.

Chapter 3 turned to money pump arguments in favour of the acyclicity of preference (in the context of certainty), which try to establish that agents with cyclical preferences stand to make a sure loss in some choice scenarios. Those arguments, I show, rely on agents being guided by their preferences in action, which is again plausible if preferences over outcomes are the standard of instrumental rationality. However, I argue that the arguments fail according to that standard. In fact, the intuition that it is bad to be money pumped relies on taking more basic desires over features of outcomes to be the standard of instrumental rationality.

Chapter 4 investigated whether money pump arguments go through if we stick with a desire-based notion of instrumental rationality. The new difficulty, I tried to show, is that according to that alternative standard, it is no longer plausible to require agents to be guided by their preferences in action under all circumstances. Reinterpreting preference as disposition to choose, I argue that money pump arguments at least give agents who desire to have stable choice dispositions a reason to have acyclical preferences.

In the context of uncertainty, as Chapter 5 shows, there is a new dimension to the question of what the standard of instrumental rationality should be taken to be. Does that standard consist in attitudes to the possible outcomes of our actions only, or also in attitudes to the uncertain prospects directly? If it is the latter, do we look to an agent's attitudes over the prospects open to her at the moment of choice, or attitudes to the prospects she faced at the beginning of a dynamic choice problem? One famous argument in favour of separability, the central requirement of expected utility theory, derives separability from, amongst other things, consequentialism and sophistication. I show that those two requirements are attractive only on distinct answers to the question of which prospects are the subjects of the attitudes that form the standard of instrumental rationality.

I go on to argue in Chapter 5 that in fact the most uncontroversial requirement of decision-making in the context of uncertainty, state-wise dominance, is only plausible if it is in fact attitudes to (features of) outcomes that form the standard of instrumental rationality. This requirement can be thought of as the requirement not to make a sure loss. I then show that we can make an instrumentalist argument to the effect that agents who violate separability may make a sure loss over time. For reasons that are parallel to those given in Chapter 4, this argument will only give agents who have a desire to have stable choice dispositions reason to adopt separable preferences. For everybody else, instrumental rationality turns out more permissive.

Chapter 6, finally, considers whether for those agents, more permissive theories like risk-weighted expected utility theory can still give structure to their decision-making under uncertainty. I argued that this is questionable: Risk-weighted expected utility theory, if it is to guard agents against sure loss, turns out to make either approximately the same recommendations as expected utility theory, or to be extremely sensitive to an agent's framing of her decision problem.

Most decision theorists, we saw in Chapter 2, appeal to no conative attitude other than preference, at least not explicitly. They appear to be committed to *preference-based instrumental rationality*, the idea that we should evaluate an agent's actions by how well they serve her preferences over the objects of choice. Ultimately, however, if we were to stick with this notion of instrumental rationality, there is very little instrumental rationality can say in favour of orthodox decision theory. It could justify why we think agents should follow *maximization*. But it could not help justify central requirements on the structure of an agent's preference, like acyclicity or separability.

Appealing to desires about features of the objects of choice, that is, features of the outcomes or prospects an agent is choosing, allows us to tell a richer story. As Chapter 3 argued, it tells a richer story about instrumental rationality more generally, by appealing to a desiderative structure that tells us *why* an agent prefers one outcome over another, not only that she does so. But it also allows us to give at least a partial defence of the core requirements of orthodox decision theory beyond *maximization*.

To the agent with non-separable or cyclical preferences, we can say the following: We can put you in a situation where, unless you are willing to act counter-preferentially, or make temporary adjustments to your preferences, you will make a sure loss. That is, you will choose, over time, such that you end up with less of something you desire, when you could have easily foregone that loss. For agents who desire not to adapt their choice behaviour to different choice contexts, the best way to avoid this sure loss is to adopt separable or acyclical preferences. Everybody else, however, is free to avoid sure loss by making those temporary adaptations to their choice behaviour. Decision theory gives good advice to those who prefer to be steady in their choice behaviour. But instrumental rationality is more permissive for the rest of us.

Bibliography

George Ainslie. *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person.* Cambridge University Press, 1992.

George Ainslie. *Breakdown of Will.* Cambridge University Press, 2001.

Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21(4):503–546, 1953.

Chrisoula Andreou. Temptation and deliberation. *Philosophical Studies*, 131(3):583–606, 2006.

Chrisoula Andreou. There are preferences and then there are preferences. In Barbara Montero and Mark D. White, editors, *Economics and the Mind.* Routledge, 2007.

Chrisoula Andreou. The real puzzle of the self-torturer: Uncovering a new dimension of instrumental rationality. *Canadian Journal of Philosophy*, 45(5-6):562–575, 2015.

Chrisoula Andreou. Cashing out the money-pump argument. *Philosophical Studies*, 173(6):1451–1455, 2016.

Frank Arntzenius and David McCarthy. Self torture and group beneficence. *Erkenntnis*, 47(1):129–44, 1997.

Kenneth Arrow. A difficulty in the concept of social welfare. *The Journal of Political Economy*, 58: 328–346, 1950.

Shlomo Benartzi and Richard Thaler. Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics*, 110(1):73–92, 1995.

Shlomo Benartzi and Richard Thaler. Myopic loss aversion and the equity premium puzzle. *Management Science*, 45(3):364–381, 1999.

Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation.* Dover Publications, 1789/2007.

Simon Blackburn. Practical tortoise raising. *Mind*, 104(416):695–711, 1995.

Luc Bovens. The value of hope. *Philosophy and Phenomenological Research*, 59(3):667–681, 1999.

Seamus Bradley. Imprecise probabilities. In Edward Zalta, editor, *Stanford Encyclopedia of Philosophy.* URL = http://plato.stanford.edu/archives/sum2015/entries/imprecise-probabilities/, summer 2015 edition, 2015.

Michael Bratman. *Intention, Plans, and Practical Reason.* Harvard University Press, 1987.

Michael Bratman. Planning and temptation. In Larry May, Marilyn Friedman, and Andy Clark, editors, *Mind and Morals: Essays on Ethics and Cognitive Science*, volume 293-310. Bradford/MIT, 1995.

Michael Bratman. Toxin, temptation, and the stability of intention. In Jules Coleman, Christopher Morris, and Gregory Kavka, editors, *Rational Commitment and Social Justice: Essays for Gregory Kavka*, pages 59–83. Cambridge University Press, 1998.

Michael Bratman. Rational planning agency. *Royal Institute of Philosophy*, 2017.

Rachael Briggs. Decision-theoretic paradoxes as voting paradoxes. *Philosophical Review*, 119(1):1–10, 2010.

Rachael Briggs. Costs of abandoning the sure-thing principle. *Canadian Journal of Philosophy*, 45(5-6): 827–840, 2016.

John Broome. *Weighing Goods.* Blackwell, 1991.

John Broome. Are intentions reasons? and how should we cope with incommensurable values? In Christopher Morris and Arthur Ripstein, editors, *Practical Rationality and Preference*, pages 98–120. Cambridge University Press, 2001.

Lara Buchak. *Risk and Rationality.* Oxford University Press, 2013.

Lara Buchak. Risk and tradeoffs. *Erkenntnis*, 79(S6):1091–1117, 2014.

Erik Carlson. Cyclical preferences and rational choice. *Theoria*, 62(1-2):144–160, 1996.

Donald Davidson, J. C. C. McKinsey, and Patrick Suppes. Outlines of a formal theory of value, i. *Philosophy of Science*, 22:140–160, 1955.

James Dreier. Rational preference: Decision theory as a theory of practical rationality. *Theory and Decision*, 40(3):249–276, 1996.

Jon Elster. *Sour Grapes: Studies in the Subversion of Rationality.* Cambridge University Press, 1983.

Lesley Fellows. Deciding how to decide: ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. *Brain*, 129:944–952, 2006.

J.K. Ford, N. Schmitt, S.L. Schechtman, S.L. Hults, and M.L. Doherty. Process tracing methods: Contributions, problems, and neglected research questions. *Organizational Behavior and Human Decision Processes*, 43:75–117, 1989.

Jerry Gaus. Reasonable utility functions and playing the cooperative way. *Critical Review of International Social and Political Philosophy*, 11:215–234, 2008.

David Gauthier. *Morals by Agreement.* Oxford University Press, 1987.

David Gauthier. Assure and threaten. *Ethics*, 104(4):690–721, 1994.

David Gauthier. Commitment and choice. In F. Farina, S. Vannucci, and F. Hahn, editors, *Ethics, Rationality, and Economic Behaviour*, pages 217–243. Oxford University Press, 1996.

David Gauthier. Resolute choice and rational deliberation: A critique and a defense. *Nous*, 31(1):1–25, 1997.

Faruk Gul and Wolfgang Pesendorfer. The case for mindless eonomics. In Andrew Caplin and Andrew Schotter, editors, *The Foundations of Positive and Normative Economics*. Oxford University Press, 2008.

Johan E. Gustafsson. A money-pump for acyclic intransitive preferences. *Dialectica*, 64(2):251–257, 2010.

Johan E. Gustafsson. Money pumps, incompleteness, and indeterminacy. *Philosophy and Phenomenological Research*, 42(1):60–72, 2016.

Walter Habenicht, Beate Scheubrein, and Ralph Scheubrein. Multiple criteria decision making. In Ulrich Derigs, editor, *Optimization and Operations Research*, volume 4, pages 257–279. EOLSS Publications, 2002.

Peter Hammond. Consequentialist foundations for expected utility. *Theory and Decision*, 25:25–78, 1988.

Jean Hampton. The failure of expected-utility theory as a theory of reason. *Economics and Philosophy*, 10(2):195–242, 1994.

Jean Hampton. Does hume have an instrumental conception of reason? *Hume Studies*, 21(1):57–74, 1995.

Daniel Hausman. Revealed preference, belief, and game theory. *Economics and Philosophy*, 16(1): 99–115, 2000.

Joseph Heath. *Following the Rules*. Oxford University Press, 2008.

Timothy Heath and Subimal Chatterjee. Asymmetric decoy effects on lower-quality versus higher-quality brands: Meta-analytic and experimental evidence. *Journal of Consumer Research*, 22:268–284, 1995.

Brian Hedden. *Reasons without Persons: Rationality, Identity, and Time*. Oxford University Press, 2015a.

Brian Hedden. Options and diachronic tragedy. *Philosophy and Phenomenological Research*, 40(2): 423–451, 2015b.

Thomas Hobbes. *Leviathan: Or the Matter, Fore, and Power of a Common-Wealth Ecclesiasticall and Civill*. Yale University Press, 2010/1651.

Richard Holton. *Willing, Wanting, Waiting*. Oxford University Press, 2009.

Joel Huber, John Payne, and Christopher Puto. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9(1):90–98, 1982.

Cynthia Huffman and Barbara Kahn. Variety for sale: Mass customization or mass confusion? *Journal of Retailing*, 74(4):491–513, 1998.

David Hume. *A Treatise of Human Nature*. Clarendon Press, 2007/1739.

Richard Jeffrey. *The Logic of Decision*. University of Chicago Press, 2nd edition, 1965/1983.

D.F. Jones and M Tamiz. *Practical Goal Programming*. Springer Books, 2010.

James Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

Daniel Kahnemann and Dan Lovallo. Timid choices and bold forecasts: A cognitive perspective on risk-taking. *Management Science*, 39(1):17–31, 1993.

Daniel Kahnemann and Amos Tversky. Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Daniel Kahnemann and Amos Tversky. Choices, values, and frames. *American Psychologist*, 39(4): 341–350, 1984.

Daniel Kahnemann, Peter Wakker, and Rakesh Sarin. Back to bentham? explorations of experienced utility. *The Quarterly Journal of Economics*, 112(2):375–405, 1997.

Gregory Kavka. The toxin puzzle. *Analysis*, 43(1):33–36, 1983.

Niko Kolodny. Why be rational? *Mind*, 114:509–563, 2005.

Niko Kolodny. How does coherence matter? *Proceedings of the Aristotelian Society*, 107:229–263, 2007.

Niko Kolodny. Why be disposed to be coherent? *Ethics*, 118(3):437–463, 2008.

Botond Köszegi and Matthew Rabin. Mistakes in choice-based welfare analysis. *American Economic Review*, 97(2):477–481, 2007.

Isaac Levi. Money pumps and diachronic books. *Proceedings of the Philosophy of Science Association*, 3:S235–S247, 2002.

David Lewis. Desire as belief. *Mind*, 97:323–332, 1988.

Sarah Lichtenstein and Paul Slovic, editors. *The Construction of Preference*. Cambridge University Press, 2006.

Graham Loomes and Robert Sugden. Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, 92(368):805–824, 1982.

Graham Loomes and Robert Sugden. Disappointment and dynamic consistency in choice under uncertainty. *The Review of Economic Studies*, 53(2):271–282, 1986.

Graham Loomes, Chris Starmer, and Robert Sugden. Observing violations of transitivity by experimental methods. *Econometrica*, 59(2):425–439, 1991.

Lola Lopes. Decision making in the short run. *Journal of Experimental Psychology: Human Perception and Performance*, 9:377–385, 1981.

Lola Lopes. When time is of the essence: averaging, aspiration, and the short run. *Journal of Experimental Psychology*, 65(3):179–189, 1996.

Mark Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27(4):1622–1668, 1989.

Patrick Maher. *Betting on Theories.* Cambridge University Press, 1993.

Edward McClennen. *Rationality and Dynamic Choice: Foundational Explorations.* Cambridge University Press, 1990.

Edward McClennen. Rationality and rules. In Peter Danielson, editor, *Modeling Rationality, Morality, and Evolution*, pages 13–40. Oxford University Press, 1998.

John Stuart Mill. *Utilitarianism.* Oxford University Press, 1861/1998.

Christopher Morris and Arthur Ripstein. Practical reason and preference. In Christopher Morris and Arthur Ripstein, editors, *Practical Rationality and Preference.* Cambridge University Press, 2001.

Donald Morrison. On the consistency of preferences in allais' paradox. *Behavioral Science*, 12(5):373–383, 1967.

Robert Nozick. *The Nature of Rationality.* Princeton University Press, 1993.

Derek Parfit. *Reasons and Persons.* Oxford University Press, 1984.

L.A. Paul. *Transformative Experience.* Oxford University Press, 2015a.

Sarah Paul. Doxastic self-control. *American Philosophical Quarterly*, 52:145–158, 2015b.

B. Peleg and M. Yaari. On the existence of a consistent course of actions when tastes are changing. *Review of Economic Studies*, 40(3):391–401, 1973.

Richard Pettigrew. Risk, rationality, and expected utility theory. *Canadian Journal of Philosophy*, 45 (5-6):798–826, 2015.

Philip Pettit. Decision theory and folk psychology. In Michael Bacharach and Susan Hurley, editors, *Foundations of Decision Theory: Issues and Advances*, pages 147–175. Blackwell, 1991.

Plato. *Protagoras.* Cambridge University Press, 2008.

John L. Pollock. *Thinking about Acting: Logical Foundations for Rational Decision-Making.* Oxford University Press, 2006.

Douglas W. Portmore. *Commonsense Consequentialism: Wherein Morality Meets Rationality.* Oxford University Press, 2011.

John Quiggin. A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4): 323–343, 1982.

Warren Quinn. The puzzle of the self-torturer. *Philosophical Studies*, 59(1), 1990.

Matthew Rabin. Risk aversion and expected utility: a calibration theorem. *Econometrica*, 68(1281-1292), 2000.

Matthew Rabin and Richard Thaler. Anomalies: risk aversion. *Journal of Economic Perspectives*, 15: 219–232, 2001.

Wlodek Rabinowicz. Money pump with foresight. In M. Almeida, editor, *Imperceptible Harms and Benefits*, pages 123–154. Kluwer, 2000.

Wlodek Rabinowicz. A centipede for intransitive preferences. *Studia Logica*, 67:167–178, 2001.

Wlodek Rabinowicz. Safeguards of a disunified mind. *Inquiry*, 57(3):356–383, 2014.

Diana Raffman. Indiscriminability and phenomenal continua. *Philosophical Perspectives*, 26:309–322, 2012.

Frank P. Ramsey. Truth and probability. In R.B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, pages 52–94. Routledge, 1928/1950.

Julian Reiss. *Philosophy of Economics*. Routledge, 2013.

Stephen Ross. Adding risks: Samuelson's fallacy of large numbers revisited. *Journal of Financial and Quantitative Analysis*, 34(3):323–339, 1999.

Paul Samuelson. Risk and uncertainty: a fallacy of large numbers. *Scientia*, 98(108-113), 1963.

Leonard Savage. *The Foundations of Statistics*. Wiley, 1954.

Frederic Schick. Dutch bookies and money pumps. *Journal of Philosophy*, 83(2):112–119, 1986.

Ulrich Schmidt. Alternatives to expected utility theory: Formal theories. In Salvador Barbera, Peter Hammond, and Christian Seidl, editors, *Handbook of Utility Theory*, pages 757–837. Kluwer, 2004.

Thomas Schwartz. Rationality and the myth of the maximum. *Nous*, 6(97-117), 1972.

Teddy Seidenfeld. Decision theory without "independence" or without "ordering". *Economics and Philosophy*, 4:267–290, 1988.

Itamar Simonson. Choice based on reasons: the case of attraction and compromise effects. *Journal of Consumer Research*, 16:158–174, 1989.

Katie Steele. What are the minimal requirements of rational choice? arguments from the sequential-decision setting. *Theory and Decision*, 68(4):463–487, 2010.

H. Orri Stefansson and Richard Bradley. How valuable are chances? *Philosophy of Science*, 82(4): 602–625, 2015.

H. Orri Stefansson and Richard Bradley. What is risk aversion? *British Journal for the Philosophy of Science*, forthcoming.

Sarah Stroud. Weakness of will. In Edward Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014.

Robert Sugden. Alternatives to expected utility theory: Foundations. In Salvador Barbera, Peter Hammond, and Christian Seidl, editors, *Handbook of Utility Theory*, pages 685–755. Kluwer, 2004.

Larry Temkin. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning.* Oxford University Press, 2012.

Sergio Tenenbaum. Minimalism about intention. *Inquiry*, 57(3):384–411, 2014a.

Sergio Tenenbaum. The perils of earnest consequentializing. *Philosophy and Phenomenological Research*, 88(1):233–240, 2014b.

Sergio Tenenbaum. Reconsidering intentions. *Nous*, 50(2), 2016.

Sergio Tenenbaum and Diana Raffman. Vague projects and the puzzle of the self-torturer. *Ethics*, 123 (1):86–112, 2012.

Michael G. Titelbaum and Matthew Kopec. The uniqueness thesis. *Philosophy Compass*, 11(4):189–200, 2016.

E. Triantaphyllou. *Multi-Criteria Decision Making: A Comparative Study.* Kluwer, 2000.

Amos Tversky. Elimination by aspects: a theory of choice. *Psychological Review*, 79(4):281–299, 1972.

David Velleman. The story of rational action. In *The Possibility of Practical Reason.* Oxford University Press, 1993/2000.

Bruno Verbeek. Consequentialism, rationality and the relevant description of outcomes. *Economics and Philosophy*, 17(02):181–205, 2001.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, 1944.

Alex Voorhoeve and Ken Binmore. Transitivity, the sorites paradox, and similarity-based decision-making. *Erkenntnis*, 64(1):101–114, 2006.

Peter Wakker. Nonexpected utility as aversion to information. *Journal of Behavioral Decision Making*, 1:169–175, 1988.

Michael Weber. The resilience of the allais paradox. *Ethics*, 109(1):94–118, 1998.

Paul Weirich. Expected utility and risk. *British Journal for the Philosophy of Science*, 37:419–442, 1986.

Roger White. Epistemic permissiveness. *Philosophical Perspectives*, 19(1):445–459, 2005.

Bernard Williams. Internal and external reasons. In Ross Harrison, editor, *Rational Action*, pages 101–13. Cambridge University Press, 1979.

Ulrich Witt. From sensory to positivist utilitarianism and back – the rehabilitation of naturalistic conjectures in the theory of demand. *Papers on Economics and Evolution*, 2005(07), 2005.