Title
Rationality and Expected Utility

Permalink
https://escholarship.org/uc/item/0bg2x07c

Author
Gee, Max

Publication Date
2015

Peer reviewed|Thesis/dissertation
Abstract

Rationality and Expected Utility

by

Maxwell Gee

Doctor of Philosophy in Philosophy

University of California, Berkeley

Professor Niko Kolodny, Co-Chair

Associate Professor Lara Buchak, Co-Chair

We commonly make a distinction between what we simply tend to do and what we would have done had we undergone an ideal reasoning process—or, in other words, what we would have done if we were perfectly rational. Formal decision theories, like Expected Utility Theory or Risk-Weighted Expected Utility Theory, have been used to model the considerations that govern rational behavior.

But questions arise when we try to articulate what this kind of modeling amounts to. Firstly, it is not clear how the components of the formal model correspond to real-world psychological or physical facts that ground judgments about what we ought to do. Secondly, there is a great deal of debate surrounding what an accurate model of rationality would look like. Theorists disagree about how much flexibility a rational agent has in weighing the risk of a loss against the value of potential gains, for example.

The goal of this project is to provide an interpretation of Expected Utility Theory whereby it explicates or represents the pressure that fundamentally governs how human agents ought to behave. That means both articulating how the components of the formal model correspond to real-world facts, and defending Expected Utility Theory against alternative formal models of rationality.
# Contents

1 Introduction

1.1 The Expected Utility Theorem and a Substantive Notion of Rationality

1.1.1 Past Attempts at Answering These Questions

1.1.2 Sen and the Problem of Action Explanation

1.2 Project Thesis and Outline

1.2.1 Chapter 1: Teleological Functionalism

1.2.2 Chapter 2: The Steps in Practical Reasoning

1.2.3 Chapter 3: Providing an EUM Interpretation

1.2.4 Chapter 4: Coherentism

1.2.5 Chapter 5: Defending the Independence Axiom

2 Teleological Functionalism

2.1 Introduction

2.1.1 Chapter Thesis and Plan

2.2 The Deontic Normative Relation

2.2.1 Challenges for a Theory of Deontic Normative Pressure

2.2.2 4 Success Conditions on a Theory of Deontic Normative Pressure

2.3 Teleological Functionalism

2.3.1 Teleological Functionalism Defined

2.3.2 Teleological Functionalism and Causal Explanation

2.3.3 Two Objections to Teleological Functionalism

2.3.4 Teleological Functionalism and Deontic Normativity: our Four Success Conditions

2.4 Foot and Biological Norms

2.5 Conclusions

3 The Steps in Practical Reasoning

3.1 Introduction

i
6.8.6 A Working Reply: Evolved Heuristics, the Discovery of Risk, and Extrapolating Intuitions to Principles

6.9 Conclusions

6.9.1 On This Chapter

6.9.2 Regarding the Overall Project

7 Appendix

7.1 Introduction

7.2 Chapter 3

7.2.1 Theorem 1: Linearity

7.2.2 Theorem 2.1: That for any outcomes, \( \omega_A \) and \( \omega_B \), that \( v(\phi) \) is in compliance with the above axioms implies that a tradeoff rate ((1)-(3) for some \( r/s \)) exists between \( \omega_A \) and \( \omega_B \).

7.2.3 Theorem 2.2: That the tradeoff rate \( r/s \) between \( \omega_A \) and \( \omega_B \) is fixed. That is, if there exists a tradeoff rate \( x/y \) between \( \omega_A \) and \( \omega_B \) then \( x/y = r/s \).

7.2.4 Theorem 3: That if \( l_1 \succ l_2 \) then there exist two partitions \( \Phi_{l_1} \) and \( \Phi_{l_2} \), each with \( n \) members, such that for all \( i \) from 1 to \( n \), \( v(\phi_{l_1i}) > v(\phi_{l_2i}) \).

7.2.5 Continuity from the Definition of \( v(\phi) \)

7.2.6 Independence from the Definition of \( v(\phi) \)

7.3 Chapter 5

7.3.1 Mission to Mars

7.3.2 Antibiotics

7.3.3 Riflery Competition
Chapter 1

Introduction

1.1 The Expected Utility Theorem and a Substantive Notion of Rationality

This project is about the applicability of Expected Utility Theory to rational human decision making. By ‘rational’ I do not mean representability by a complete and transitive preference ordering (although that may ultimately be a consequence of rationality). We make a common distinction between what we simply tend to do and what we would have done had we undergone an ideal reasoning process. For example, I might fail to accurately calculate the expected income from an investment, or I might be overcome by a desire for cake even though I would prefer to eat healthy. In cases like this there is a sense in which am not behaving ideally, that I am not as I ought to be—that I am getting in my own way. This sense of ideal decision making is what I mean by rationality. So another way of putting the goal of this project is to provide an interpretation of Expected Utility Theory whereby it explicates or represents the structure of deontic normative pressure—the pressure that fundamentally governs how human agents ought to behave. This description might not be perfectly clear baldly stated, so I want to make it a bit clearer by highlighting what exactly is involved in meeting this goal and why one would want to do it or expect it could be done.

We will take Expected Utility Theory to refer, first and foremost, to a structure that is almost purely formal. This structure can be characterized by the Expected Utility Theorem, which is just the idea that if a ranking obeys certain axioms, then it will be possible to assign numbers called utilities to each object being ranked such that vector space mixtures of those objects have utility numbers according to the
following rule when $N$ is the number of objects

\[
\text{(Utility of the mixture)} = \sum_{n=1}^{N} (\text{Utility of object } n) \ast (\text{Weight of object } n \text{ in the vector}).
\]

(Mas-Colell, Whinston, & Green 1995) constructs the formal theory as follows:

- a set $\Omega$ of outcomes that are the possible consequences of actions. We will assume that the set $\Omega$ is finite and we will index the members by $n = 1, \ldots, N$.

- a set $L$ of possible lotteries, or lists ($l_1 = (p_1, \ldots, p_N)$), that assign a probability $p_n \in (0, 1)$ to each member of $\Omega$. $\sum_{n=1}^{N} p_n = 1$

- a binary preference relation $\succsim$. Where if ($l_1 \succsim l_2$) and it is not the case that ($l_2 \succsim l_1$), then we say ($l_1 \succ l_2$).

- **Completeness** - for any two lotteries $l_1, l_2 \in L$, $l_1 \succsim l_2$, $l_2 \succsim l_1$, or both ($l_1 \sim l_2$).

- **Transitivity** - for any three lotteries $l_1, l_2, l_3 \in L$, if $l_1 \succsim l_2$ and $l_2 \succsim l_3$, then $l_1 \succsim l_3$.

- **Continuity** - for any three lotteries $l_1, l_2, l_3 \in L$, if $l_1 \succsim l_2 \succsim l_3$, then there exists a $p \in [0, 1]$ such that the agent is indifferent between a $p$ probability of $l_1$ coupled with a $1-p$ probability of $l_3$ and probability 1 of $l_2$. In our notation: $(p, l_1; (1-p), l_3) \sim (1-p, l_2; p, l_3)$.

- **Independence** - for any three lotteries $l_1, l_2, l_3 \in L$, if $l_1 \succsim l_2$ and it is not the case that $l_2 \succsim l_1$, then for any $p \in [0, 1)$, $((1-p), l_1; p, l_3) \succ ((1-p), l_2; p, l_3)$.

- **Expected Utility Theorem** - If agents’ preferences over lotteries obey the four preceding axioms, then it is possible to construct a function $U : L \rightarrow \mathbb{R}$ such that $l_1 \succsim l_2$ if and only if $U(l_1) \geq U(l_2)$. Furthermore, $U(\cdot)$ has expected utility form, meaning we can assign a real number $u$ to each $\omega \in \Omega$ such that for any lottery $l = (p_1, \ldots, p_N)$, $U(l) = u_1p_1 + \ldots + u_Np_N$. Therefore, for any two lotteries $l_1 = (p_1, \ldots, p_N), l_2 = (p'_1, \ldots, p'_N) \in L$, $l_1 \succsim l_2$ if and only if $\sum_{n=1}^{N} u_np_n \geq \sum_{n=1}^{N} u_np'_n$.

  1. The utility function is unique up to positive affine transformation. So if $U$ is a utility function representing an agent’s preferences, and $U' = aU + x$, then $U'$ will also represent those preferences.
I say this theory is *almost* purely formal because in any interpretation of the formal model that goes by the name Expected Utility Theory, these objects are always interpreted as the outcomes of actions, this vector space of combinations is always interpreted as a space of possible lotteries—or actions that carry a certain probability of bringing about different outcomes—and the axioms represent an agent relative to the standard of maximizing expected utility. We will call these common interpretive elements the *essential interpretation*—the bare bones interpretation without which the formal structure is hard to understand as being about ‘expected utility’ at all.

Coupled with the essential interpretation, I will refer to this model as expected utility maximization (EUM). As such, EUM has been used to model human decision, bacterial reproduction, tiger mating behavior, and animal foraging, just to name a few. The minimal components in the essential interpretation give EUM this wide applicability. In this project I am interested in EUM as a particular type of model of human behavior.

EUM provides a standard against which actions can be measured. Assuming we can assess an agent’s utilities, we can judge whether his actions are in compliance with EUM. But anyone can create any rule, or standard, and compare human behavior to it. This wouldn’t necessarily be a worthwhile task. For example, I could create a rule *chooses things never before chosen* and corresponding formal model, and then represent human actions relative to that rule. A model like this might be part of a story of how adventurous someone is, but there will be a lot left out with respect to the structure of human decision making. A model like this will not tell me, for example, *why* an agent chooses the way she does, or how she will choose between options she has already experienced in the past.

EUM is an interesting model for human behavior because of its promise to explicate or display worthwhile information about that behavior. EUM has been used to predict, plan, evaluate, and explain human action. For example, if people tend to maximize expected utility, and you can discover an agent’s utilities, you can tell what he will probably do when facing choices involving outcomes with those utilities.

Each intended use of Expected Utility Theory, whether it is to predict, plan, evaluate, or explain human action, will need to come along with a story of why the formal model is at all relevant to that intended use. In the case of prediction, this story will regard why that kind of behavior has followed, and will plausibly continue to follow, this pattern. In the case of planning, we will need to know why maximizing expected utility is a good idea. Coupled with these more in-depth descriptions, EUM can take on a *topic specific interpretations* that allow it to contribute to a complete explanation of why behavior follows EUM, or ought to. Specifically, in order to have a full explanation we need to ask three questions corresponding to the three parts of
the theory.

1. Why the outcomes are differentiated the way they are?

2. What governs the utility assignments (what are preference orderings about)?

3. Why does, or why ought, the agent in question maximize utility (comply with the axioms)?

1.1.1 Past Attempts at Answering These Questions

Older definitions of utility (Mill, Bentham, Sidgwick, Edgeworth), take what is typically called a realist notion of utility. On theories like these, utility tracks some property of the world, or our own psychologies, at which human action aims. This takes a step towards providing a story for why agents do, or ought to, maximize expected utility. If utility is happiness, for example, then it seems plausible to expect people to aim at it—it also might seem that there is a good justification for aiming at it.

The problem with views like this is that it isn’t clear what property could do all the things we want utility to do. Take, for example, a classical approach to interpreting utility for deontic normative application—the idea that utility is just the agent’s own happiness. An objection to this view is that it intuitively seems that people ought to make choices that do not obviously maximize their own happiness, like helping others in need, for example. So we might try again and equate utility with the aggregate happiness of all people. But in addition to returning potentially counter-intuitive results, this theory doesn’t have a clear grip on motivation. One’s own happiness seems to have an undeniable appeal to oneself that the happiness of others just doesn’t have because what is good about happiness, first and foremost, is the experiencing of it. If I ought to maximize the happiness of others, or if it is reasonable to expect that I will, there would need to be a component of the topic specific interpretation explaining why.

Topic specific interpretations allow us to find different notions of utility for the different purposes to which EUM is put. For example, one might look at neurobiological evidence to justify EUM as a predictive model. On an approach like this, one would develop a topic specific interpretation where EUM was about neurobiological facts. This would be a different completed theory from a completed EUM theory of what people ought to do. We might try to justify a theory like this by looking for similarities between EUM and plausible moral philosophies—a very different approach than for justifying the predictive model. On a theory set to this purpose,
utility would then correspond to whatever property in the world or in our minds we ought to maximize, rather than one we merely tend to maximize.

Both of these approaches face contemporary and historical objections. This project is about addressing contemporary problems. But it is instructive to look at how theorists wrestled with related issues in the past in order to get clear about what the contemporary concern really is. Originally, the resistance to realist notions of utility was motivated by limitations in psychology and neuroscience, and a distrust of the intuitions-based method of moral philosophy. Without a method of investigating the micro-structures of the brain and their impact on choice, theorists in the early 20th century wanted to maintain scientific respectability by relying exclusively upon testable phenomena. But this creates puzzles for EUM explanation. If utility is supposed to represent a property that we ought to maximize, for example, there is no obvious way to scientifically investigate this property. After all, what experiment could possibly reveal that you ought to do something?

There are two alternatives to realist notions of utility. One is to be an anti-realist and think that utility is just an artifact of the theory representing observable phenomena\(^1\). Another is to be agnostic about whether utility refers to anything and just assert the claim that people’s behavior can be informatively modeled using utilities anyway. Revealed Preference Theory is a good example of both of these approaches to the utility concept. RPT gives meaning to the preference relation, and utilities thereby, using the following additions (Mas-Colell, Whinston, & Green 1995, 13):

\begin{itemize}
  \item a correspondence \( C(B) \), where \( B \subseteq L \) and \( B \) is nonempty. Applied to actions \( C(B) \) is the nonempty set of lotteries the agent might select when faced with the alternatives available in \( B \). \( C(B) \subseteq B \) for all \( B \subseteq L \).
  \item The Weak Axiom of Revealed Preference (WARP): Let \( N, M \subseteq L \). If \( x, y \in N \cap M \), \( x \in C(N) \) and \( y \in C(M) \) then \( x \in C(M) \).
  \item The Induced Choice Rule: Let \( B \subseteq L \). Define \( \succeq \) such that for all \( a_i \in B \), \( a^* \succeq a_i \) just in case \( a^* \in C(B) \).
  \item Revealed Preference Theorem - WARP holds for all choices among all possible menus \( N, M \subseteq L \) of at least three members, if and only if \( \succeq \) is complete and transitive.
\end{itemize}

\(^1\)The distinction between realism and anti-realism will turn out to be a matter of degree. For now we can just compare the two poles of that gradient.
The presentation of Revealed Preference Theory is vague between two readings—one is epistemic and one is ontological. This happens because the “just in case” in the definition of the Induced Choice Rule is vague between merely correlating preferences (and thereby utilities) with patterns of choices and defining preferences in terms of patterns in choices. On a purely epistemic reading, RPT just tells you how to discover an agent’s utilities, and is agnostic about what utilities represent. On an ontological reading, RPT tells you that utilities are about patterns in choice behavior.

The tendency to take the ontological reading of Revealed Preference theory is supported by things its creator explicitly says. Paul Samuelson developed RPT in his *Foundations of Economic Analysis*. He was reacting to widespread concerns, like those above, that economics based on a realist notion of utility was untenable. This includes understandings of utility in terms of some impersonal metaphysical good, as well as in terms of psychological states. In assessing the state of economics at the time, he says that not only has there been a move away from the utilitarianism of Bentham, Sidgwick, and Edgeworth, but “concomitantly, there has been a shift in emphasis away from the physiological and psychological hedonistic, introspective aspects of utility” (Samuelson 1947, 91).

Following Samuelson, some theorists working with RPT took themselves to be providing explanations of human behavior by using a model that, through the definitions of its component parts, only made claims about other behavior. This was supposed to be an improvement over the use of the dubious concepts in realist utility: “...the new formulation is scientifically more respectable [since] if an individual’s behavior is consistent, then it must be possible to explain that behavior without reference to anything other than behavior” (I. M. D. Little, 1949, 90).

### 1.1.2 Sen and the Problem of Action Explanation

The problem with the Revealed Preference approach on either reading is that it doesn’t explain why we should expect human action to conform with EUM or why it ought to. This is because “preference” in the RPT model doesn’t track anything that explains the influences on decision making.

On this interpretation the use of the word “preference” in revealed preference would appear to represent an elaborate pun. In saying that x is revealed preferred to y, it would not be asserted that x is preferred to

---

2 A similar issues comes up in the work of Ramsey and Savage which I will discuss in detail in chapter 4.
y in the usual sense of the word “preferred”. A redefinition of the expression “preference” is, of course, possible, but it is then legitimate to ask what does “consistency” of behavior stand for and on what basis are the required consistency conditions chosen... Preferring x to y is inconsistent with preferring y to x, but if it is asserted that choice has nothing to do with preference, then choosing x rather than y in one case and y rather than x in another need not necessarily be at all inconsistent. What makes them look inconsistent is precisely the peep into the head of the consumer, the avoidance of which is alleged to be the aim of the revealed preference approach (Sen 1973, 243).

Sen’s language here is vague between several possible objections that all fall under the same heading. ‘Inconsistent’ can mean several things, some of which are compatible with RPT and some of which aren’t. If ‘inconsistent’ just means ‘deviant from a defined rule’, then this example is inconsistent with the axioms. Creating rules solely for the purpose of measurement—that is, using EUM in a way where it is nothing more than a set of rules relative to which actions can be described—I will call a non-deontic normative use of EUM. Non-deontic norms are just rules, or standards, that have nothing to say about the mechanism by which human beings actually choose what to do, or what governs what they ought to do. They are just rules that can provide a metric for the comparison of actions with one another. On an ontological reading, or an epistemological reading that is agnostic about the nature of utility, RPT is perfectly capable of providing non-deontic norms.

The problem with RPT is if it is used for prediction, planning, explanation, or deontic normative evaluation without being supplemented by a realist notion of utility. With respect to planning and explanation, ‘inconsistent’ can mean inconsistent with the normal causes of human action. One might think that the only reason EUM would be a capable predictor of human action is if it was reasonably isomorphic with the structures that actually cause human action. This would provide a kind of explanation for why we are justified in expecting human agents to maximize expected utility. But it would require that preferences and utilities are, at least in part, about the stuff inside the head that actually makes people do the things they do. That means the theory requires a realist conception of utility.

Setting aside the question of predicting human behavior, deontic normative evaluation is the evaluation of action relative to what an agent ought to do. It is the standard in light of which an action is a good thing to choose, or “makes good sense”. This is the way we use EUM when we use it to plan our own actions or evaluate choices as good or bad decisions. The deontic normative use of RPT seems to be partly what Sen has in mind. In (Sen 1973) he goes on to discuss the concept of
welfare in welfare economics. Of the different notions of inconsistency here, welfare is most closely related to the idea that maximizing expected utility makes good sense, not just that it is predictable or in compliance with a non-deontic standard.

‘Inconsistent’ in this context means inconsistent with the standards of good decision making. Because people often do things that they shouldn’t do, this sense of ‘inconsistent’ can come apart from the motivational sense of ‘inconsistent’ from the previous paragraph. If RPT is agnostic about the referent of utility and preference, or if it equates utility and preference with choice behavior, it fails to provide a reason to think complying with EUM is what we ought to do. Suppose I selected \( x, y \in N \cap M, x \in C(N) \) and \( y \in C(M) \) and then decided against \( x \in C(M) \). If someone objected “But you chose \( x, y \in N \cap M, x \in C(N) \) and \( y \in C(M) \)” I could justifiably respond “So what?” I am not constrained by a matter of physical necessity to obey WARP. Additionally, my past choice behavior does not, by itself, without any further description or explanation, provide an reason to act in any particular way now. So if EUM is given this kind of behavioristic interpretation, we are left without a complete story as to why I ought to conform.

What this suggests is that both deontic-normative and predictive EUM modeling require a realist notion of the components of the model if they are supposed to provide an explanation of the tendency, or deontic normative pressure, to behave in a way that is compliant with the axioms.

But if we can’t remain agnostic about the notion of utility and remain justified in a conviction that EUM provides worthwhile standards, what is the notion of utility that is relevant? We know that choice behavior is insufficient. But does utility represent a psychological state, a property of the mind-independent external world, or perhaps some combination of both? In the deontic normative context, what facts could possibly ground a justifiable objective judgment that someone ought to do something? In trying to provide a realist conception of utility that serves this purpose, we run headlong into the concerns that led Samuelson to depart from realist notions of utility in the first place. In focusing on deontic normative uses of EUM in this project, I am addressing whether we can provide a scientifically and philosophically respectable notion of utility that grounds ‘ought’ judgments and justifies the use of EUM in a substantive theory of practical rationality.

1.2 Project Thesis and Outline

This project aims to provide a topic specific interpretation of EUM for deontic normative purposes. This would make the model relevant for planning actions, or assessing agents’ behavior as a good or bad idea. Another way of putting this is that we
are looking at EUM as a potential explication of a substantive notion of practical rationality—rationality as the faculty that governs my ability to do what it makes good sense to do.

Part of a natural view of human agency is that people can engage in a process whereby they determine what it makes sense to do and then actually do it by way of their reasoning and will. If this is true, then there are two conditions that need to be met in order to provide a topic specific interpretation for deontic normative EUM. Firstly, the common sense notion of practical rationality needs to find support and precise exposition. Secondly, this account of practical rationality, whatever it turns out to be, needs to resemble expected utility maximization such that EUM is a good representation of the structure of practical reasoning.

There is a common sense or naïve interpretation that people give to EUM. It’s the interpretation one might give if you took the names of all the components of the theory literally, or perhaps if you came to the theory for the first time without knowing about its behavioristic roots. This is the idea that preference orderings are just psychological preferences, outcomes are the aims of my actions, and lotteries capture my beliefs about the likelihood that a choice brings about a particular outcome. A view like this has some superficial plausibility in light of how intuitive it is. It seems like my preferences, or desires, come in strengths that might be reflected in utilities. And it seems like EUM might describe the way I can be most effective at satisfying my desires. Being as effective as possible at satisfying my desires is at least a candidate for what it could mean for an action to make good sense. This creates an obvious resemblance with the quasi-Humean picture of practical rationality—one where what I ought to do is governed by the preferences/desires that I have. The task of providing a deontic normative interpretation of EUM would then be to spell out this connection more explicitly.

However, since (Sen 1973) many have argued, on different grounds, that this view cannot work. These objections address both the neurological and psychological facts that would ground reasoning along the lines of EUM, and the possibly counterintuitive results of thinking that what I ought to do is maximize the expected satisfaction of my preferences, whatever they are. These contemporary objections drive the five chapters that follow. I will argue that none of these objections actually is decisive. Indeed we can provide a well-justified topic specific interpretation of EUM for deontic normative purposes that is close to this common sense notion of rationality, and isn’t inconsistent with anything we know about the scientific picture of the human mind. So the folk-psychological and introspective plausibility of the view, coupled with the fact that there are no troubling objections to it, will leave us in a position at the end of this project where we have a working possible interpretation of EUM
as a theory of quasi-Humean rationality.

By the end of this project I hope to have presented and defended a view with the following answers our three questions about topic specific interpretations:

1. Outcomes are differentiated on the basis of the content of the objects of an agent’s desires.

2. Utility assignments are about beliefs an agent has about the degree to which he ought to pursue lotteries. These beliefs are a response to an agent’s desires.

3. An agent rationally judges that he ought to maximize expected utility as a logical consequence of these basic ‘ought’ beliefs. This form of rational judgment is all there is to the fact that the agent ought to do something.

1.2.1 Chapter 1: Teleological Functionalism

Deontic normative decision theory needs to answer the same three questions that other topic specific interpretations need to answer—how do we differentiate outcomes, why are utilities assigned in a particular way and what are they about, and why does a rational agent maximize utility? But normative decision theory faces an additional hurdle. EUM as a deontic normative representation presents perfectly rational action. But human agents are predictably irrational. If our goal is to model, plan action for, and give advice to, actual human agents, then we need a sense in which an EUM model can both display perfect rationality, and be about imperfect irrational agents.

The way to do this is for the model to display how an agent is aspiring to be—why he is getting in his own way, or making a mistake whenever he deviates from that model. It would be sufficient for our purposes to construct a model that meets this standard. It would tell the agent how he should plan his own actions, what advice we should give him, and how to evaluate his actions as well-chosen or poorly-chosen.

So a plausible starting point for answering the three questions above is to give a description of the nature of deontic normative pressure. The question I address in Chapter 1 is “what kind of fact could govern all cases of good decision making?” I need a fact that is scientifically respectable in the sense that we have good empirical justification for believing in it, and that is inescapably attractive for all agents we would want to call rational. I will settle on a position, Teleological Functionalism, that claims that we have an introspective access to a first-personal property of our minds whereby we aim at different outcomes. This property serves as our candidate ground for ‘ought’ facts.
1.2.2 Chapter 2: The Steps in Practical Reasoning

Chapter 2 is about the specific Teleological Functionalist aims that our mental states have. I will demonstrate how beliefs, desires, and intentions come together to create an intuitive notion of practical rationality as *instrumental* rationality—rationality as taking the means to our ends. Because each of the mental states involved in creating instrumental rationality has its own kind of aim property, Chapter 2 explicates the compositional nature of our goals in practical reasoning.

The key component at each step in practical reasoning is the universalizability of that step. That is, there must be an explanation why the agents we would naturally want to call rational would all see the appeal of taking that next step towards doing what they ought to do. So even if we don’t always live up to the standards of rationality, we can come to understand the sense in which we’ve *messed up* when we don’t.

At this level of granularity, philosophical considerations motivate components of the theory that go beyond the common sense notion of practical rationality. My hope is that the position is close enough to common practice to retain a minor presumption of truth.

1.2.3 Chapter 3: Providing an EUM Interpretation

Chapter 3 interprets EUM in terms of the picture of rationality presented in chapter 2 and demonstrates how this interpretation avoids objections grounded in the fact that neurobiological forces governing decision making predictably cause EUM violations.

Meacham and Weisberg describe this problem as the tension between normative and characterizational representationalism. I will discuss versions of this problem from various authors (e.g., Korsgaard, Davidson, Hulse, Read, and Schroeder). The basic idea is that characterizing the way human beings actually are is in tension with characterizing how they ought to be. The question is, how can modeling an ideal agent say anything about pressure on human beings—messy imperfect collections of causally governed biology?

This objection breaks down into two more precise types. The first is that the picture of human psychology presented in this chapter is not anything like the causal reality of the brain. But if EUM does not accurately characterize the way our brains are, then it cannot tell us anything about pressure on us to be any particular way. The second version of this objection is that if EUM was an accurate causal characterization, it would be impossible for us to act otherwise than as perfect EUM agents. But this would imply that we are always exactly as we ought to be, and would
therefore contradict the common sense notion that we make mistakes of practical reasoning all the time.

Chapter 3 employs the structure of Teleological Functionalism, and the distinction between first-personal and third-personal mental properties, to address these objections.

1.2.4 Chapter 4: Coherentism

By chapter 4 my position has been completely presented. Here I move to address objections from alternative positions. This chapter compares instrumental rationality to rationality through coherence as such. The difference between these views has been described in many different ways, but the basic question is what makes it the case that there is deontic normative pressure towards reasoning in a particular way. Coherentism says that mental states and reasoning are the way they are because it is coherent—that is, there is some magnetism towards mental states that have a certain kind of harmony or consistency with one another. Instrumentalism is the view that normative pressure derives from facts about what is effective at accomplishing what we are trying to do.

The difference can be seen in two approaches to how one might think about decision theoretic norms. Suppose I prefer A to B and yet, in a choice between them, I pick B. It is compatible with coherence that I do one of two things—I could revise my preference so that I prefer B to A, or I could stop choosing B over A. Both are paths to coherence. But instrumental rationality might give me only one option. What is required is that I do the thing that is most effective at satisfying my preferences given what they are. This means that my only option is to stop choosing B.

Practical reasoning may be made up of coherence norms, instrumental norms, some combination of both, and/or something else entirely. I investigate the work of four authors who think that decision theory represents coherence norms—Lewis, Davidson, Dreier, and Armendt. Ultimately I conclude that rationality based in coherence could never explain universalizability—that, insofar as he understands the norm, a rational agent will always appreciate the pressure of instrumental norms but could reject any norm of coherence as such.

1.2.5 Chapter 5: Defending the Independence Axiom

Taken as a constraint on good decision making, Independence is the most controversial of the EUM axioms. For some, taking the most effective means to satisfying our
desires seems to permit Independence violations in certain cases. Take, for example, the case of Maurice Allias. He finds it intuitive that an agent could prefer $L_1$ to $L_2$ and $L_4$ to $L_3$ in the following two choice problems.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2-11</th>
<th>12-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$</td>
<td>$0$</td>
<td>$5M$</td>
<td>$0$</td>
</tr>
<tr>
<td>$L_2$</td>
<td>$1M$</td>
<td>$1M$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2-11</th>
<th>12-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_3$</td>
<td>$0$</td>
<td>$5M$</td>
<td>$1M$</td>
</tr>
<tr>
<td>$L_4$</td>
<td>$1M$</td>
<td>$1M$</td>
<td>$1M$</td>
</tr>
</tbody>
</table>

In *Risk and Rationality*, Lara Buchak suggests that Allais might have the preferences he does because he reasons in the following way:

[in the choice between $L_1$ and $L_2$] He reasons that the minimum he stands to walk away with is the same either way, and there is not much difference in his chances of winning some money. So, since $L_1$ yields much higher winnings at only slightly lower odds, he decides he would rather have $L_1$...

[between $L_3$ and $L_4$] He reasons that the minimum amount that he stands to win in $L_4$ is a great deal higher than the minimum amount he stands to win in $L_3$, and that although $L_3$ comes with the possibility of much higher winnings, this fact is not enough to offset the possibility of ending up with nothing. So he decides he would rather have $L_4$ (12).

Chapter 5 compares EUM with Buchak’s competitor, Risk-Weighted Expected Utility Theory (REU). REU rejects Independence and can explain Allais’s preferences. In this chapter I defend the Independence axiom by bringing up some strange consequences of violating it. In the literature there are many such examples. What is different about what I do here is that I investigate the implications of Independence violation for concepts and practices on the periphery of individual decision making—things like group decision and inferences about efficacy. My hope is that, through focusing on the broader implications of Independence violation, we can distance ourselves from decision making heuristics that might have colored theorists’ intuitions about these cases in the past.
Chapter 2

Teleological Functionalism

2.1 Introduction

The substantive notion of practical rationality—the one that I claim justifies modeling agents in terms of expected utility maximization—is the capacity to determine what you ought to do and then actually do it by way of reasoning and the will. This suggests that there are what I will call deontic normative relationships between certain types of things.

A deontic normative relationship is an ‘ought’ relationship between a state of an agent and the state of the world around him, or a state of an agent and another state of that agent. So, for example, if there is suffering nearby, it might be the case that I ought to stop it. The relationship between the suffering and my action would be a deontic normative relationship. Or maybe I want to go to college. In that case, I ought to study—another deontic normative relationship.

Deontic normative relationships might not just be between initial states and ultimate actions, but there might be intermediary steps in between. Suppose, for example, that what I wanted at the most basic level was wealth, and I knew that college was the best path to wealth. Then maybe my desire for wealth implies that I ought to desire to go to college which implies that I ought to study.

A theory of practical reasoning needs to answer two questions. Firstly, what is the nature of a deontic normative relationship? How does it have the force that it seems to have in planning action, giving advice, evaluating agents, etc.? Secondly, what are the steps, or rules governing the steps, between different deontic normative relationships? Another way of putting this is what are the relata of these relationships?

This chapter is about that first question. The following chapter is about the
2.1.1 Chapter Thesis and Plan

Deontic normative relationships are puzzling for at least two reasons. The first is that they are not straightforward consequences of properties and forces that we have independent reason to believe exist. The second is that deontic normative relationships, and the ‘ought’ and reasons statements that they entail, have an inconsistent relationship with action, so it is not obvious how they factor into action explanations.

But before I can describe these puzzles in depth, I need to give a more detailed description of deontic normative relations and the role they play in a theory of rationality. To that end, in the next section I’ll give a more in-depth description of what a deontic normative relation is like and the challenges involved in trying to give a precise account of how they work. This will lead to four success conditions that a theory of the nature of deontic normative pressure needs to meet.

Section 3 contains my original contribution to the discussion. I will argue that there is a common way of talking about mental states, Teleological Functionalism, that can give an account of deontic normative relations compatible with all four success conditions.

Section 4 will contrast this view with an alternative theory of the nature of normative relations from Foot and Thompson.

Section 5 will outline work still to be done and the implications of these conclusions for the rest of the project.

2.2 The Deontic Normative Relation

Part of rationality is determining what you ought to do through reasoning. This means that there must be a standard, or norm, in light of which you can judge something to be a good or bad thing to do. But not all norms are about the pressure to be rational. A norm is just a standard, different from a mere trend, in light of which something can be judged adequate or inadequate, functioning or malfunctioning, good or bad, etc. Normativity is just measurement in relation to a standard. So, for example, the rule that men walk on the street-side of the sidewalk when walking with a woman defines a norm, or standard. But this is different from a statistical trend, as nowadays most men don’t actually follow that rule, or even see it as implying reason for them to comply. In a case like this, I will say that there is normativity,
but it is not *deontic*, or reason-implying\(^\text{12}\). That is, there is a standard in light of which a man can be compared, but it is not a standard that provides an undeniable reason for action.

A case of a deontic norm might be something like ‘If you want to play tennis, get a racket’. This norm implies a reason that is inescapable in the sense that the sidewalk norm is not. While both norms are, in a sense, hypothetical, they are contingent in structurally different ways. In the sidewalk case, I can disavow the norm itself—I may decide to walk on the building-side of the sidewalk. In the case of the tennis norm, the force of the norm may be contingent upon my desire to play tennis, but once I want to play tennis, the norm seems inescapable. The sidewalk norm also has this kind of contingency—if I don’t want to go for a walk with my date, then I don’t need to abide by this norm. But the tennis norm doesn’t have the additional contingency where I can outright reject it like I can the sidewalk norm. If I want, all things considered, to play tennis and yet refuse to get a racket, there is something deeply confused or irrational about my position. What we are looking for in theories of deontic normativity is whether and how there could be any norms that meet this kind of inescapability.

Normative decision theory, as intended by some economists, is non-deontic. That is, while the axioms of expected utility maximization provide a standard against which the behavior of agents can be measured, these theorists are agnostic about whether a person *ought to obey* these axioms. Gary Becker’s *A Treatise on the Family* is an example of a work that assumes agents can be meaningfully represented by classical utility theory without implying a claim that it is how people ‘ought’ to be. Additionally, there are philosophers, like Foot and Thompson, who mean to give theories of deontic normativity in the sense that I mean it, but, in my opinion, only provide accounts of non-deontic norms\(^\text{3}\). For these reasons I want to separate these two concepts.

The previous examples looked at cases of norms, deontic and not, without statistical trends. Most people might not give to charity, or walk on the right side of the

---

\(^1\) Other examples of non-deontic norms might include having two arms, or having wisdom teeth. Both of these properties could be used to define a standard against which individual cases could be measured. Furthermore, they could both continue to define a standard independently of trends in a population—if almost everyone lost an arm, for example. But neither standard necessarily provides a reason to do anything. That is, there is no reason why the non-compliant should take themselves as being under pressure to comply with the standard.

\(^2\) I do not mean ‘deontic’ in the sense that the rule is the most fundamental *source*, in some sense, of reasons. I just mean deontic as action guiding, or as implying a reason that the agent would endorse if they understood it, whatever the ultimate nature or source of that reason might be.

\(^3\) See (Foot 2001), or (Thompson 2003)
sidewalk. In either case the statistical trend is irrelevant to whether there is a norm. The same goes for cases where there are trends without norms. For example, most people might live in the eastern hemisphere. This doesn’t mean that it is wrong or inadequate to live in the western hemisphere. However, I could define a non-deontic norm—living in the eastern hemisphere—against which people could be measured regardless of where most people happen to live.

All of this suggests a concept which will be ubiquitous throughout the discussion that follows which is *deontic normative pressure*. Deontic normative pressure is the pressure to be in compliance with deontic norms. It might be the case that pressure to be in compliance with a deontic norm can be overridden by other considerations. But it does seem, at first glance, that there is a pressure like this that theories of rationality are trying to capture\(^4\).

### 2.2.1 Challenges for a Theory of Deontic Normative Pressure

**Challenge 1: ‘Ought’ Facts and Independent Support**

To reiterate, practical reasoning is a process by which rational agents determine what they ought to do and do it. This means that there must be some part of that process by which rational agents can come to judgments of the form “I ought to do X”. These judgments have similar qualities to judgments about objective facts. We might debate them as right or wrong, or critique someone for failing to appreciate that they ought to do something. And as we saw above, they have a kind of inescapability. This seems to suggest that deontic normative relationships are grounded in some sort of objective ‘ought’ facts that ‘ought’ judgments are attempts to track.

But the problem is that we do not have any independent justification for claiming the existence of ‘ought’ facts.

I think the most straightforward way to understand this problem is to contrast it with cases of strong justification. Take, for example, the case of electrons. I might have a theory about the nature of atoms that requires electrons in order to work. But until I have some kind of experimental or observational evidence for electrons themselves, my theory is unsupported. The idea is that the fact that my theory

---

\(^4\)This pressure is something like what philosophers might call the force of pro tanto reasons. The benefit of the term ‘normative pressure’, I think, is that it is specifically about the motivational component of a reason, rather than the fact a person ought respond to (like suffering nearby, for example). Additionally, the term is agnostic about any distinction between reasons and the pressure of rationality that arises in some work (see for example (Kolodny 2005)).
requires electrons in order to work is not itself a good reason to think there are electrons—I need something else.

This epistemic constraint isn’t specific to this case. Rather, it is an instance of a general principle that applies because progressing without it could lead us into error. Take the case of ancient scientists trying to explain fire. They might come up with a theory where a spirit is responsible for bringing fire whenever two sticks are rubbed together. But this wouldn’t be a good reason in itself to believe in the spirit. This is because there are other possible explanations for why the fire happens, and the fact that a particular theory requires the spirit does not rule out those alternative possibilities. Really, what these ancient scientists need is a way to rule out other possible explanations, either by finding independent empirical support for the existence of the spirit, or through finding a method, perhaps conceptual, to demonstrate the impossibility of other options.

So the problem, in its most distilled form, is that the fact that a theory requires the existence of an entity is not a reason to believe in the existence of that entity as long as there are other live possible explanations. So the ‘independent support’ we need in order to vindicate the theory and believe in the requisite entity is either ruling out other the other possibilities, or finding independent evidence for the existence of the entity.

This same consideration applies to theories of rationality and the purported facts that they require. Any theory of rationality will require facts that make it the case that “if you want to play tennis, you ought to get a racket”, for example. But that a theory requires it is not enough to justify the claim that ‘ought’ facts exist. Notice that there are other possible explanations of our observation of general agreement about the tennis racket: maybe people are under a systematic illusion when they make ‘ought’ judgments. If this were true, then the phenomena that our theory attempts to explain are actually caused by something else. And it isn’t at all obvious how we could rule out other possibilities like this. In fact, these alternative positions are very much a live option in the debate about the nature of ‘ought’ judgments.

Without an obvious path to ruling out alternatives on conceptual grounds, it seems like our best option for providing support for the existence of ‘ought’ facts is finding empirical evidence. But here’s the puzzle: none of the facts for which there is straightforward empirical support are either directly about what we ought to do or conceptually entail anything about what we ought to do. Thinking of the direct case, we might ask “What type of observation would empirically confirm that murder is wrong?” But all the observations we might make are about a categorically different

\[5\text{Take, for example, expressivists like Hare who think that ‘ought’ judgments just express positive attitudes towards things, making the appearance of objectivity in these judgments a kind of illusion.}\]
subject matter. We might observe that there was a living thing, and then there wasn’t a living thing, that certain chemical reactions were taking place, and then weren’t, or that a person participated in a society, and then didn’t. But there is a conceptual difference between judging that these things are the case and judging that I ought to do anything. This holds for any property that I can directly observe or measure. No object or property that belongs to the lexicon of fundamental scientific properties or forces corresponds to our required ‘ought’ fact.

But even if we cannot directly pick up on ‘ought’ facts, perhaps we can infer their existence. This does not seem to work either. None of the fundamental properties and forces we can detect seem to imply anything about ‘ought’ facts. Facts about what we ought to do are not the right kind of thing to be the product of a chemical reaction (or, at least, do not present themselves as such). Nor are ‘ought’ facts a logical consequence of facts about what is the case. So I could never come to a conclusion about what ought to be by way of a logical deduction from judgments about what is, nor could I infer that an ‘ought’ fact will result as a causal consequence of some other non-deontic facts. So even if I could give a complete description of all the atoms in the universe and their interactions with one another, I could fail to explain why anyone ought to do anything. Therefore, all the empirical support I could get about natural objects and their properties leaves me without empirical support for the existence of ‘ought’ facts. So if we are looking for justification for beliefs in ‘ought’ facts, it isn’t clear that we can lend support to those views by measuring anything or by making inferences about what we can measure.

In summary, a theory of objective ‘ought’ judgments seems to require facts that we can’t empirically support, and we can’t immediately rule out alternative possible theories that do not require these facts.

Challenge 2: Causal Inconsistency and Rational Explanation

A rational explanation is an explanation of action that articulates why an agent did what he did in a light that rationalizes that action, or presents it as an expression of practical rationality. For example, I might explain why Bob went to the store by mentioning that he wanted frozen pizza and he knew that there was frozen pizza at the store. This explanation lets you know what Bob found appealing about going to the store—it lets you know how his action fits a deontic norm, “take the means to your ends”, or something like this. This contrasts with an explanation of rational action, which might give a neurobiological, for example, explanation of why an action occurred without articulating the sense in which that action was ideal or well-chosen.

(Davidson 1963) argues that rational explanations must be causal explanations.
He says this because we need to be able to distinguish between reasons for which an action was performed and reasons that merely stand in favor of performing an action. Suppose it is cold out and my jacket is fashionable. I have two reasons for wearing my jacket. But if I don’t care about fashion, then temperature is the only reason that actually explains why I did what I did. The difference here seems to be the causal impact of the reason on my action.

But the causal force of an action is not one and the same as its appeal in reasoning. Irrationality is a state where someone is subject to the demands of rationality, but fails to conform to deontic normative pressure. Take a case of weakness of will. A student might need to study, recognize that fact, desire to study (in some sense of ‘desire’), and yet fail to do so. In this case, the reason was present and action didn’t follow. So the appeal that the irrational agent appreciated about his reason to study came apart from the efficacy of that reason in generating action.

Furthermore, notice that we now have a challenge of rational explanation. In Bob’s case, citing a reason is sufficient to give an explanation. But in the student’s case, citing a reason is insufficient to give an explanation. The challenge we face is how to give a sufficient causal explanation by citing reasons given that reasons are sometimes causally inefficacious. This is just a special case of a general challenge associated with causal explanation. I might explain why a match lit by citing the fact that I struck it. But if it is wet it will not light even if I strike it. So the fact that I struck the match is an insufficient explanation. In any causal explanation, there are additional facts that we leave out that would be part of a more complete explanation of why the event occurred. I aim to give a description of what those additional facts are in light of which citing a reason can be an explanation like “I struck the match.”

2.2.2 4 Success Conditions on a Theory of Deontic Normative Pressure

The broad description of practical rationality and the challenges above suggest four standards I want to meet in presenting a theory of the nature of deontic normative relationships.

Present a Causal Story of Rational Explanation

Deontic normative relationships have to be compatible with being causal explanations for why things happen even though reasons, by themselves, are insufficient explanations. “I did it because it was the right thing to do,” even though sometimes
I don’t do the right thing, for example. So we need a coherent causal story from the beginning of the practical reasoning process to the end—a story that fills the gaps akin to how striking a match is just part of the story of what gets the match to light. That means, firstly, that whatever property in the world leads to judgments about what one ought to do needs to be accessible through sense perception and cognition. Otherwise, there would be no causal story of how those facts led to action. Additionally, whatever appeal or pressure an agent experiences from the perception of a reason can stand as a cause, but will fail to do so in cases of irrationality. So we need to understand how citing reasons can explain action when, at times, reasons retain their appeal for an agent without being expressed in action.

**Explain the Appeal of Deontic Norms**

Both rational and irrational agents must be under some sort of pressure to be rational. This pressure is something these agents have access to through reasoning, which is a conscious process. So there must be something appealing, *from their point of view*, about being rational.

This appeal defines rational norms and separates them from mere trends. Actions that are not in compliance with this appeal, or pressure, are irrational. Because it creates a standard relative to which actions can be measured, it is normative. And because it is appealing, it is deontic.

A good heuristic for testing whether we have met this condition for a particular agent is the ‘So What?’ test. If an agent is non-compliant with a purported norm of rationality, we should be able to bring up that norm and the substantive cost of his non-compliance should be apparent to him. If he just says ‘So what? I don’t care about that standard’, then there are two possibilities: either he is non-rational, a term I will define shortly, or we have failed to identify a true deontic normative constraint.

**Independent Empirical Support**

Whatever entities this theory requires as the grounds of deontic normative relationships (‘ought’ facts), considerations governing theory justification requires that we find support for their existence independently of the need for them in the theory. For example, we might justify our belief in ought facts as logical, or conceptual, consequences of other commitments we already hold. Alternatively, we could justify our belief in ought facts because we have empirical access to them in some form.
Extensional Adequacy

Theories of rationality naturally create three categories of living things: rational, irrational, and non-rational. Rational and irrational agents are subject to rational norms. That means that both appreciate the appeal of deontic normative pressure. Agents are rational to the extent that they act in accordance with that pressure, and irrational otherwise. Non-rational living things are not subject to rational norms. Any living thing that cannot reason about what they ought to do is non-rational. Squirrels, amoebas, etc., might fall into this category.

Extensional adequacy is the idea that our theory should imply that the right things to fall into each of these three categories. So, by and large, human beings should come out as either irrational or rational (at a particular moment, with respect to a particular decision). Amoebas should be non-rational.

The fact that an agent responds ‘So What?’ to a purported deontic norm might mean that he is a non-rational agent. Imagine an agent who feels no desire to help others and responds ‘So What?’ to related deontic norms. If a theory makes caring about others a fundamental deontic norm—one which is not the result or implication of other types of norms—then this agent gets categorized as non-rational on that theory. On this theory, and with respect to these sorts of norms, irrationality is when one cares about people and fails to follow through in action.

But it would be better if the theory did not return the result that common practice is systematically misguided. While that is a defeasible consideration, it does at least provide a kind of evidence against the truth of a theory. So if a theory returns the result that an agent who is otherwise capable, and who we would naturally want to hold accountable to rational and moral norms, can coherently respond ‘So What?’ to a norm of that theory, then this counts as evidence against that theory.

A consideration that will come up later in this project is whether there can be multiple different types of rationality. There may be non-rational agents, and then different sets of rational and irrational agents based upon the nature of the norms that they respond to. Throughout this project I will leave this as a live possibility, being content to argue that human beings seem, in general, to abide by a rationality that resembles expected utility maximization.

2.3 Teleological Functionalism

Practical rationality is about how people respond to different facts about themselves and their environment. This means that a search for the nature of deontic normative pressure needs to focus, at least in part, on mental states and how they facilitate
rational responses.

If we could show that people ought to have certain mental states as responses to things, or ought to respond to mental states in certain ways, then we could develop a map of practical reasoning. But what are the states of the mind/brain that are relevant for a map like this? One way to define these states would be in terms of causal roles. For example, we could define ‘beliefs’ as mental states caused by perceptions in such and such cases, and that cause other mental states and actions in some list of cases. This list would presumably be difficult to create. However, causal definitions could potentially be conceptually sound even if we couldn’t, at a particular time, or with our current understanding, accurately map those relationships.

But there is another reason why we need to use a different set of considerations to differentiate the roles of mental states for our purposes: people deviate from what they ought to do all the time. Presumably, when they do this, they did it in compliance with causal laws. This means that if there was normative pressure to act some other way, then that pressure must have deviated from causal pressure. So if we define mental states in terms of deontic normative roles, then those roles cannot reduce in any straightforward way to causal laws about what people actually do, or tend to do. So we need a sufficiently robust and independent alternative that explains why this pressure arises from the kinds of sensitivities intrinsic to these mental states that can act as an alternative to a causal definition.

Additionally, notice that the definition of a deontic normative relation does not require a causal connection anyway. Rather, it requires an ‘ought’ relationship that may or may not result in an agent actually living up to the standard. The demand is that rational and irrational agents see the appeal in reasoning a particular way, which is just a restriction on how an agent thinks about what to do from the first person point of view. While it will be necessary that in the brain there is some causal story or other that underlies and facilitates this appeal, and the possibility of following through in action, the first-personal appeal of a method of reasoning could provide an alternative set of rules in light of which we could define mental states.

We can see how this might look by contrasting it with a traditional causal functionalist account of beliefs. Our causal functionalist defines beliefs as above—a long list of causal roles. But I will argue that we can define beliefs in terms of first-personal properties instead.

It is natural to think that there is a certain kind of goal-directed activity people are engaged in when they believe things. That is, what I am trying to do in believing is represent for myself the way the world is. This aim is an irreducibly first-person property of my mental states. It is something I can become aware of, and something that is conceptually distinct from any combination or third-personal facts about me.
or my brain. We can use this first-personal aim-property to define ‘beliefs’: Beliefs are for the purpose of representing facts about the way the world is for deliberation. This definition is not incompatible with there being a causal story surrounding belief. There will be causal facts that underlie the manifestation of these first-personal aims, we just aren’t defining beliefs in terms of them. And importantly, a definition in terms of a first-personal aim will provide a standard in light of which we can explain why an agent finds a particular standard of behavior or reasoning appealing—it is an expression of the aim of the mental state(s) under discussion.

I will call these first-personal aim-properties of my mind teleological functions. Beliefs are all instances of a similar type of aim, and, as such, belong to the same teleological functional category.

That mental states have properties with exclusively first-personal ontology, and that they have properties akin to teleological functions, are not new ideas. With respect to first person ontology, there are plenty of canonical examples. They operate by way of demonstrating that explanations using purely third-person accessible facts are limited because they do not capture the first-person character of experience. The phenomenological, intentional, and (I will argue) teleological facts about mental life cannot be captured simply by talking about actions or neurons and how they appear to interact from the third-person perspective. Suppose that we have a complete third-personal description of the universe, physics and neuroscience included. Someone who knew all those things, still might not know certain facts about mental life. If she has never seen them, she might not know what it is like to see colors (Jackson 1982). Or she might not know what it is like to be a being with different sensory modalities, like echolocation (Nagel 1974). The total of third-personal facts seems to leave out how certain things, like books, brain states, thermometers, are about other things (Searle 1980, Searle 1983). But most importantly for our purposes, the third person perspective leaves out the sense in which an action was done on purpose or for a purpose—how an action is something that the agent chose to do for reasons, and not something that just happened to him. The fact that the third-person perspective leaves out what is done on purpose will give us an avenue to describing how it leaves out what is done justifiably or rationally.

My argument for the inadequacy of the third-person perspective for describing what is done on purpose is twofold. First I want to argue that there is a feeling associated with doing something on purpose—there is something it is like. Secondly I will argue that, without recourse to first-personal facts, it is impossible to delineate neurological states on the basis of whether they are on purpose, and as such one must start with the first-personal property and then move from there to find a neurological correlate. The reason what is done on purpose is especially relevant to this project is
that responding to rational norms is something that we ought to purposefully do if we notice that we are not in compliance. Whatever subject matter we are responding to in making purposeful choices—that is, whatever it is that norms of rationality are about—must be first-personally accessible. It must be possible to give a first-personal account of rational norms given the fact that those norms inspire action by way of our first-personal access to their content.

then it must be possible to give a first-personal account of rational norms as well.

With respect to our first task, John Searle does the same thing in (Searle 1983). Searle cites two experiments, one where a person tries and fails to raise his arm, and another where a person’s hand is moved by stimulating his brain, and the patient felt as though he did not move his hand on purpose, rather, that the experimenter did it. The first case comes from (James 1950) where a patient is anesthetized and told to close his eyes. His arm is then tied down and he is told to raise his arm. After attempting to raise his arm the patient is told to open his eyes and, to his surprise, his arm has not gone up. Now, this is not to show that there is no neurological correlate of the feeling of trying to raise your arm. If we track the effects of the anesthetic, we may even be able to find it. But certainly the patient doesn’t have access to his neurological states as such, and he did have a sense that his arm went up. That sense, since his arm didn’t go up and he didn’t see his arm, probably comes from some sort of feeling associated with his action. This also seems introspectively plausible.

But maybe this is inconclusive. Maybe what he felt was not trying to raise his arm but rather that his arm was raised, and this feeling simply wasn’t veridical. But there is another experiment by Wilder Penfield that gets closer to the issue:

When I have caused a conscious patient to move his hand by applying an electrode to the motor cortex of one hemisphere I have often asked him about it. Invariably his response was: “I didn’t do that. You did.”

When I caused him to vocalize, he said, “I didn’t make that sound. You pulled it out of me.” (Penfield 1950 as quoted by Searle 1983)

Once again, this says nothing about whether there is a neurological correlate of the first-personal sense of acting on purpose. But it certainly must be the case that what the patient was aware of was not the absence of a neurological state as such, but rather the absence of a feeling that normally attends his moving his hand or vocalizing.

So even if there is a neurological correlate of doing something on purpose, it is accompanied by a type of first-personal feeling. Now suppose there is a neurological correlate of doing something on purpose. Respectful of the canonical arguments
above, it would be impossible to determine what that feeling was like from third-personal information alone\textsuperscript{6}. Furthermore, if it were possible to derive a concept of what is done on purpose relative to what simply happens to an agent without using first-personal facts, it is not at all obvious how one would go about doing that. Take the case of a cellist playing a complex piece and making a mistake. From the neurological point of view, everything that happened will be determined by the laws of physics as the laws of physics say it is to be determined—there would be no mistake relative to that picture of what is happening. Even if there was a neurological correlate of what the agent was trying to do and that was uncharacteristically discordant with some system that generates the agent’s actual performance, and this uncharacteristic discordance could account for the mistake\textsuperscript{7}, even that divergence would have been determined in accordance with physical laws. It is not clear how we could derive the idea that one of those systems was \textit{supposed to be} like the other as opposed to simply \textit{tended to be} like the other. This difference in how actions and action generation are supposed to be versus simply how they tend to be is intimately related with what is done on purpose and is most naturally understood from the first person point of view. As was mentioned above, there is a feeling associated with doing something on purpose\textsuperscript{8}. Furthermore, the concept of doing something on purpose itself seems specific to our first-personal mental life. Even if this were incorrect, and we could find some third-personal candidate for doing something on purpose, it could not be distinguished as such without reference to first-personal facts.

### 2.3.1 Teleological Functionalism Defined

Teleological Functionalism is a method for defining mental state terms whereby we distinguish mental states in terms of their \textit{teleological functional roles}, which are the

\textsuperscript{6}This is why these neurological experiments require the subject to be awake. In order to find the neural correlate of a conscious experience, we need a report from the first-personal experience of the subject. We can’t just look at the neurons and tell what that feels like.

\textsuperscript{7}This is a different meaning of the word ‘mistake’ than I meant to imply above when talking about normative considerations. The cellist is trying to play one thing and then fails to do what she was trying to do. A mistake with respect to the norms of rationality usually isn’t like this—it’s not that I tried to maximize expected utility and failed, but that I didn’t take it seriously enough, didn’t notice, or something like this. Even though these two types of mistakes are different, there is a sense in which they are both in the same category of failures to do what an agent commits himself to do.

\textsuperscript{8}The feeling of doing something \textit{on purpose} is different from the teleological ‘aim property’ that I define below (you could intentionally do something in conflict with the teleological properties of your mental states). All I want to do here is highlight how there are essential phenomena of action explanation that are left out by the third-person point of view.
roles they are supposed to play in practical deliberation. These roles are not purely
artifactual constructions for easily representing human behavior, rather, they capture
a very real first-personal ontological property—the teleology of our mental processes,
or, in other words, trying to do something. Teleological Functionalism is supported
by, and depends upon, the observation that there are essentially goal-directed mental
activities that can serve as plausible referents for the terms ‘belief’, ‘desire’, and
‘intention’.

Like the case of belief, there seems to be something I am up to—some aim of
my mental process—when desiring and intending—there are components of what it
is to do these things that include the felt character and the sense in which they
are for some end that is impossible to capture using just facts about what neurons
are active and what an agent happens to do. Each of these mental state categories
will correspond to a teleological functional role category. Individual mental states
will fulfill their function in different ways. For example, the function of belief may
be to represent information, and my belief about my favorite type of apple is for
representing that information.

The notion that has historically been called upon to capture the idea of teleologi-
cal functional roles in theorizing about mental states is direction of fit: “It is the aim
of a belief to be true, and to the extent that a belief is true, it succeeds... Desires, on
the other hand, are not supposed to represent how the world is, but how we would
like it to be... in the case of the desire it is, so to speak, the responsibility of the world
to fit the content of the desire” (Searle 2004, p. 118). If we rephrase the figurative
language here, it seems plausible that Searle is saying that, as far as planning action
goes, the desire makes it your responsibility to make the world fit the content of the
desire\(^9\). This quote explicates the idea of a direction of fit. Beliefs have what Searle
calls a ‘word-to-world’ (or mind-to-world) direction of fit because it is their aim to
match the world. Desires have a ‘world-to-word’ direction of fit because it is their
aim to get you to make the world match the content of the desire.

For Searle (and I agree), direction of fit is an ontologically first-personal fact. It
is lost if you attempt a purely third-personal neurological or behavioral description
of an agent (Searle 2004, p. 81).

Direction of fit is a great first step in capturing the first-personal directedness of
mental states. The shortcoming of direction of fit is that it leaves out how mental
states meet their conditions of satisfaction by way of functional interactions with

\(^9\)In the next chapter, when I talk about the formal notion of desire, it will be clear that with
respect to the kinds of desires we are interested in, it is more accurate to say that the desire
‘constitutes your endorsement of the idea that it is your responsibility’ more so than ‘makes it your
responsibility’.
one another. For example, above, Searle says that “to the extent that a belief is true, it succeeds.” But being true isn’t enough for the success of a belief. That belief needs to present that information in a usable form for the operation of other mental states. Without that feature, it is not even clear that you have a belief at all. This part of the aim and definition of beliefs is left out by direction of fit. While defining mental states in terms of direction of fit leaves you without an explanation of teleological functional roles or the relationships between mental states essential to practical reasoning, if we start with TF roles, we can use those to understand the steps in rationality, and why a mental state has a certain direction of fit and conditions of satisfaction. The next chapter is on the specific details of teleological functional roles and the steps in practical reasoning. For now I just want to clarify what Teleological Functionalism is and what it is not.

In settling on functional definition we are situating ourselves within a long tradition of folk psychological description of mental state terms (Lewis 1994 is an example). But there are two ways I want to distinguish what I am doing here from what has traditionally been called functionalism in the philosophy of mind. The first is that functionalists like Lewis take causal forces to be the functional connections that define the relationships between mental states. For example, that the definition of ‘desire’ would include causing intentions in certain conditions. But causal relationships are just one of a set of properties on the basis of which we can delineate mental states.

The second way that what I am doing here is going to deviate from traditional functionalism is that while I endorse the idea that functional definition is the best way to define mental state terms, that is itself completely separate from a full discussion of the nature of mental states themselves. Another, perhaps less esoteric way to put this point is that a functional definition may be perfectly capable of carving up a space into clearly delineated objects without giving a complete explanation of the properties used in that delineation. For example, while Teleological Functionalism is perfectly good at carving up first-personal space, it will not tell us how the first-personal properties of conscious experience arise out of third-personal neurological facts. So it uses first-personal facts in order to delineate mental states, but it does not explain what first-personal facts are—it just employs them. Traditional functionalism is supposed to provide a story where there is nothing left to explain after you provide a the map of causal interactions between mental states. So the essence, or nature, of the state is completely explicated by the theory. In this way, the ‘functionalist’ theory that I am going to endorse is not a complete theory of the nature of mental states (even if it assumes that their nature has first-personal and third-personal components), it is just a way of talking about them and their relations to one another.
Teleological Functionalism may seem esoteric at first but a subset of everyday folk psychological terms work this way. This is the nature of our language when we say things like “the essential defining characteristic of beliefs is that they store information for use by conscious processes”. Variants of teleological functionalism are actually a very common methods of differentiating the parts various systems. Other examples include the terms ‘powertrain’ in car manufacturing, ‘capital’ in economics, or ‘memory’ in computer science. In each case, the category is defined in virtue of what the object is supposed to be able to do—not the causal properties it actually has.

The nature of the ‘supposed’ in each case might be a bit different. The ‘supposed’ applied to material goods gives us the conceptual resources to explain what it means for something to malfunction. In these cases, the teleology (the ‘supposed’) is derived—the objects to which those terms refer have a purpose only in virtue of we use them. We attribute teleology to them relative to our purposes, they do not have purposes in themselves. In the case of the teleology of mental states, that teleology is intrinsic to the system, as expressed in the discussion of direction of fit and first-personal properties, and not the result of the purposes of something external to the mind. And just like derived teleology gives us the conceptual resources for understanding when a piece of equipment is malfunctioning, mental teleology gives us the conceptual resources for understanding when a mental state has a kind of malfunction—the kind whereby it contributes to an instance of irrationality.

Teleological Functionalist terms can be broken down into two parts:

1. **The definition.** A TF general term is defined in terms of a category of functional roles in a system. We then distinguish TF particulars in terms of their differing in their TF functional role within that category. For example, a simple TF framework could differentiate beliefs and desires from one another on the grounds that beliefs are for the purpose of storing information about the world and desires are not. Further, a belief about radishes can differ from a belief about triangles in that they are both for the purpose of storing different information.

2. **The instantiation.** Whatever type of stuff fills a particular TF role in a system. In the case of the mind, it would be the first-personal mental properties and third-personal neurological properties that are responsible for the performance of a particular teleological role. Instantiations are often very relevant to the details of TF objects. For example, we know that beliefs cannot have unlimited

---

10This is the parallel of the same distinction with regard to intrinsic and derived intentionality.
information storage capacity. This isn’t in virtue of their essential definition, it is a consequence of coincidental features of their instantiation in us.

None of this is to suggest that the causal powers that an object actually has are completely irrelevant to having teleological functional roles. If I put a rock in place of a car starter, it wouldn’t perform any of the functions of a starter regardless of the aims I had when I put it there. The teleological properties that something has are linked, as we will see in the next section, to the causal powers that it could express when functioning properly. Nevertheless, it is the teleological functional roles, and not the causal powers themselves, that allow us to explain the sense in which it can malfunction at all.

One might object to my decision to associate purposes with mental state tokens, and not just with mental state types. For example, one might claim that individual beliefs don’t have different purposes. Maybe I have only one purpose for beliefs in general—something like “aim at believing true things”. On a theory like this beliefs can be differentiated on the basis of their content, and not their aim. This is an interesting question that would require a deeper investigation than I can pursue here. I will say that I am not sure whether this is a substantive, or merely terminological, dispute. I might talk in terms of constitutive aims, or overall aims, and they are not obviously mutually exclusive. A student may aim to finish a paper, and aim to graduate. Similarly, in being in a state where I aim at representing something about radishes might just be constitutive of having an aim to believe true things.

One point to emphasize in order to avoid confusion about this position is that mental states, like beliefs, are not sub-personal parts of a human being. It isn’t like I have a belief as a part of my brain, and that part has aims. Rather, a belief is a state of me as an entire organism—one that is intimately related to neurological activity to be sure, but one which is most easily grasped by thinking about what I am doing, not what a part of me is doing. So, when I believe that radishes are red, I aim to represent a fact about radishes. Beliefs are just a category of my aims. ‘Aims’, in this sense, are not all-consuming. I may have other aims simultaneously—some that are in different categories and some that may directly conflict with a given mental states. Phrased this way, it sounds a bit less strange to suggest that there are first-person teleological properties.

2.3.2 Teleological Functionalism and Causal Explanation

Given the fact that I have selected folk psychological terms that are defined in terms of their first-person teleological character, one might have the following worry about the truthfulness of a folk psychological explanation using these terms. (Davidson
1963 and 1970) argue that all rational explanation must be causal explanation. This is because in order to differentiate between the reason for which an action was performed and a reason which merely counts in favor of performing an action, there must be some causal link between the effective reason and the action. But given the fact that I am defining mental states in terms of their first-personal properties, and nomological cause is a function of third-personal physical facts, how is it that citing mental states thus defined can contribute to the explanation of action?

Let’s take an example of rational action using beliefs and desires as defined above. In order to provide a rational explanation for why an agent performed a particular action you can cite a desire and a belief of the agent demonstrating that the agent believed that the action was a means to achieving the desire. For example, if someone asks “Why did Bob go to the store?” if it is true that Bob has the cited belief and desire then the following would be a rational explanation “Bob wanted a frozen pizza and he believed that he could purchase one at the store”. People give explanations of this kind all the time, but our question is whether an explanation like this could work if the ‘belief’ and ‘desire’ were understood as TF terms.

Even though beliefs and desires are defined in terms of their first-personal properties in our approach, they can still provide components of a causal explanation. There are two ways to see that this is so. Firstly, remember that even though TF folk psychological objects are defined in terms of their first-personal properties, they still have an instantiation in the brain with third personal properties. TF terms carve up the mind along first-personal lines, but they still refer to the entire object, first-personal properties and third-personal instantiation. So while having particular causal powers is not built into the definition explicitly, causal power is implied by the definition insofar as the teleological role requires it. How exactly this works, how exactly the relationship between the teleological properties and the third personal properties and whether the first-personal can be causal is not one I am trying to answer. My point here is only that in referencing beliefs and desires we reference states of a causally efficacious object.

The second reason to think that TF folk psychological terms can contribute to causal explanation is that TF terms in other domains contribute to causal explanations in the same way. Take the term ‘starter’ in car manufacturing. The starter is the component of the car that ignites the engine’s operation under its own power. A starter can work in many different ways, it might be hydraulic, pneumatic, or electric, for example—and within those categories there are many different varieties. However, if someone asked, “What gets the engine going?” it would be perfectly acceptable to say “An electric starter,” even though the term ‘starter’ is defined in terms of teleological functional role and not the causal powers it actually has (it
would still be a starter even if it was malfunctioning). The reason citing the electric starter counts as an explanation of the engine starting up is because it describes what is playing a role that had to be played by something. The starter is not sufficient to start the engine—it wouldn’t start without gas or many other parts—and this particular starter is not necessary to start the car, as it could have been started with a different starter. But talking about this particular starter indicates what object is playing a relevant role to starting up the engine, and saying that it is electric provides some detail that differentiates the way this particular starter plays that role from how other starters might have played that role. Citing the starter explains the activity of the engine relative to a background understanding of the way cars work. If the audience understands that background, then the explanation will be useful.

This is perfectly analogous to the case of citing a belief and a desire in the case of rational explanation. If Bob went to the store then there are certain roles that had to be played given a background understanding of the way people reason. Something had to motivate Bob—whatever plays that role we can call his desire. Saying that Bob went to the store “because he wanted to” is not informative because it just reiterates the background assumption about what desires are for. But if you include the information that it was a desire for frozen pizza, then we learn something about what it is that was playing a role that had to be played by something. Often, we don’t even cite the relevant belief for just this reason. A perfectly good response to “Why did Bob go to the store” is “Because he wanted frozen pizza”—it is part of our background understanding of people that they operate by accessing information about the world through beliefs and that they take the means to their ends. So appending our explanation of Bob’s behavior with the fact that Bob believed that going to the store was a means to frozen pizza is redundant, even if it is true. And just like our case with the car starter, any particular belief and desire are not necessary to perform a given action because they could have been performed because of other beliefs and desires, and they are not sufficient, as there may be a malfunction either in the belief and/or desire itself, or elsewhere in the system.

When we do what our beliefs, desires, and intentions recommend that we do, those actions are in fact caused by our beliefs, desires, and intentions. However, that does not mean that we do not have those states when we fail to do what those states recommend, or that citing them cannot explain why we did what we did when we are rational.

In summary, TF folk psychological explanations are perfectly good causal explanations because they illuminate some of what is going on in a system and explain an action relative to a background understanding of the way the system works. TF folk psychological objects have causal powers, they just aren’t defined in terms of them.
2.3.3 Two Objections to Teleological Functionslism

Objection 1: The Epistemology of Teleology

There are two epistemological concerns one might have at this point. The first is how we are supposed to determine the teleological functional definitions of our mental states. The second is that one might think we do not have introspective access to mental teleology at all, and as such would question whether this first-personal property actually exists.

With respect to the first question, my response will become much more clear in the second chapter when I actually work out what the teleological functional roles of the relevant mental states are. But, in outline, there are two points to note about how to determine these roles. The first point is that we are defining mental states in terms of these roles, so the roles are, in a sense, stipulative. It is not as if there is a well-defined belief object floating around in my mind that I can examine and determine what its teleological functional role is. Rather, I notice that I aim at different kinds of things. The different types of aiming that I do are used to define mental state terms\(^\text{11}\). This brings me to my second point about determining these roles, which is that it is by answering a ‘How possible’ question about aims. For example, if I aim to respond to the way the world is, it is possible to break down that aim into two parts. The first is aiming to know, or represent, or store, the state of the world, and the second is to execute an action in response. This is a purely stipulative carving up of a goal-directed function I have. But carving up things this way will help us define a mental state vocabulary that captures practical reasoning\(^\text{12}\).

The second epistemic concern is that one might think we do not have introspective access to a first-person teleology. When I introspect on my beliefs, for example, I may find their content (what I believe), but what part of that introspective process is the teleological property?

Believing in this first-person means property is motivated in part by the experiments above highlighting the first-person character of doing something on purpose, which, I think, has aiming at something as a constituent. But we can also motivate this view through reflection. Think of a case where you are resisting a desire. It is at least plausible that you are introspectively aware of the fact that the desire rec-

\(^{11}\)It can be helpful to remember that a belief is a state of my mind/brain. My mind/brain is the object. So we are stipulating the meaning of terms that will be useful for describing it.

\(^{12}\)The epistemology of teleological functional roles is very similar to the epistemology of the teleology of biological functions on Foot’s view. I will discuss this view in the next section. While the views differ in terms of where the teleology is located, the feature of breaking down teleological functions by ‘How possible’ questions is the same.
ommends deciding to do something. In other words, you are resisting ‘a temptation towards x’. This recommendation, this aiming at, of which you are introspectively aware, is the teleology of desire.

You can do something similar with beliefs. Reflect on a belief of yours that is in doubt. Your awareness of the inadequacy of this belief is dependent upon your awareness that the belief aims at something. Your awareness of this aim is the first-personal teleology of belief.

One reason people might be motivated to doubt their introspective access to this property is because they are making the mistake I mentioned above: looking for a part of their introspective experience that corresponds to a belief, and then trying to find its teleology. But because beliefs are not sub-personal parts, you can’t introspectively seek them out and notice that one aims at something and another aims at something else. Rather, you notice yourself aiming at things in different ways and to different degrees. So as long as you are aware of when you are aiming at things and what you are aiming at, you are aware of the first-personal teleological property I am talking about.

Objection 2: Circularity in the Explanation of Deontic Normative Pressure

One plausible deontic norm that teleological functionalism will ultimately need to explain is instrumental rationality. Instrumental rationality is something like a norm requiring that we take the means to our formal ends. So whatever it is that we ultimately decide we want, rationality requires that we seek out and take the means to getting it.

Teleological Functionalism seems to risk explaining instrumental rationality in terms of itself. Why should I take the means to my aims? Well, because your desires aim at getting you to take the means to your aims. The obvious question being, then why take the means to the aims of my desires?

The first thing I want to note is what I do not intend to do, which is to analyze or explain what it means to aim at something in the most basic sense, or how an aim connects us with its end. I’m not sure that anything illuminating can be said on that subject. ‘Aiming at something’ may be a basic concept corresponding to an irreducible constituent of mental life that cannot be meaningfully analyzed or explained in other terms. In any event, I cannot provide an analysis and do not mean to try. However, what I do want to do is demonstrate how instrumental rationality—understood as performing actions that are the means to my all-things-considered desires—is the result of the expression of constitutive smaller aims: the
aims of individual mental states.

As we will see in the next chapter, TF mental states interact with one another in the expression of their roles. Instrumental rationality will end up resulting from the interaction of the roles of belief, desire, and intention. This means that the aim of instrumental rationality is explained by TF mental state aims in the sense of being built up out of them.

This TF description will not reduce aims to properties that are ultimately not aims. In other words, I am not trying to reduce the deontic normative to something which is not deontic normative. Mental teleology is an irreducible first-personal ontological property, and so deontic normativity is irreducible as well. While teleological functional roles are discovered by answering ‘How possible’ questions, being a teleological functional role is more than just being an answer to such a question—it is being part of the expression of an aim.

So the relationship between instrumental rationality and the aims of TF mental states is not circular—rather, the latter constitutes the former and is irreducible to other types of properties.

With respect to the ‘So What?’ question, the capacity to have teleological functionalist desires, beliefs, intentions, implies that there are certain aims that you already endorse. Take the case of belief. Because a belief is defined in terms of an aim to represent the way the world is, if you have a belief, you already endorse the idea that you ought to represent relevant facts about the way the world is. So the question ‘why should I take the means to my end of representing the world’ doesn’t make sense—seeing yourself as having reason to pursue those means just is what having a belief amounts to. Insofar as you are engaged in believing, you are interested in taking the means to that end. A similar point will arise with respect to the notion of desire that is employed in the next chapter. The formal notion of desire is of a disposition to decide that you ought to pursue something. To have this disposition is to endorse the idea that you are inclined to satisfy this desire, at least commensurate with its strength\(^{13}\).

Giving an explanation of practical rationality requires situating it in a context that meets our four success conditions. This can be done as long as it is a consequence of the expression of the teleological functional roles of my mental states. As we will see, this does not require reducing the deontic normativity of practical rationality to non-deontic-normative facts.

\(^{13}\)The view I will ultimately construct is one where expected utility maximization is a consequence of weighing the aims of desires against one another. There may be ways of weighing the aims of different mental states—beliefs versus desires, for example—but I do not discuss that here.
2.3.4 Teleological Functionalism and Deontic Normativity: our Four Success Conditions

Teleological Functionalism provides us with the set of ‘ought’ facts that ground ‘ought’ judgments—a set of first-personal facts about the aims of our mental states. My claim, which I will make good on in the next chapter, is that teleological functional roles can be defined for our mental states such that it would be plausible to say that expressing the teleological functional roles of our mental states is what we ought to do and what is practically rational. But in addition to checking up on the consequences of teleological functionalism for the content of our ‘ought’ judgments, we need to check TF against our four success conditions for a good theory of the nature of deontic norms.

Present a Causal Story

In the subsection on TF and causal explanation I covered how TF defines mental states so that teleology can come apart from causal efficacy. But when TF mental states are functioning properly, they provide a sufficient cause for the responses they normatively require.

Explain the Appeal of Deontic Norms

As I described in my response to the objection of circularity, the teleology of my mental states is inescapable. Insofar as I have beliefs, desires, etc., I see myself as having certain types of reasons. So all norms grounded in teleological functional roles will be deontic.

Independent Empirical Support

Through the experiments highlighting the first-personal experience of doing things on purpose, and my response to Objection 1, I argued that we have introspective access to the teleology of mental states. This isn’t empirical evidence in the sense of third-personal empirical evidence. But it is empirical evidence nonetheless. It is just internal to one’s own experience.

Extensional Adequacy

All entities with teleological functionalist beliefs, desires, and intentions (as I define them) will be subject to the rational norms of expected utility maximization. This seems to be an appropriately broad net. It is not clear that we would want to hold a
being who did not have the aims of TF desires, beliefs, or intentions, to the standards of EUM. Furthermore, it seems sufficiently ambitious to claim that all those agents will be subject to EUM.

We have also carved out a nice distinction between the rational and the irrational. The rational are those who express their teleological functional roles, and the irrational are those who do not. The irrational experience pressure to be rational in that they have these aims in virtue of having TF mental states, but fail to live up to their own aims.

2.4 Foot and Biological Norms

In order to get clear about Teleological Functionalism as an account of deontic normative relations, it can be useful to contrast it with a similar view. Philippa Foot holds a view where we can understand deontic normativity on the basis of what I will call biological norms. Biological norms are norms such that ‘ought’ facts that ground the norm are found in our conception of human beings as a certain kind of organism containing systems that have different biological functions.

These biological functions have purposes in virtue of answering ‘How possible’ questions about an organism’s normal lifecycle. For example, members of a species have digestive systems that make it possible for them to nourish themselves. Bees dance to communicate the location of pollen. Human beings have cognitive systems that allow for them to reason. These are all functions constitutive of living the kind of life particular to the species in question. These functions can serve as standards for defining norms. In the same way that there is something “malfunctioning” about a bee that doesn’t dance, there is something wrong with a human that doesn’t obey instrumental rationality.

Standards defined this way successfully create norms in light of which we could judge living things or functions to be adequate or inadequate. Furthermore, these standards are are different from mere trends. The majority of members of a species could fail to digest correctly, for example, and that would not be contrary to the notion that their digestive system has a purpose. We would just say that their digestive system usually does not fulfill its purpose. Additionally, in answering ‘how possible’ questions, biological norms are meant to explain what is good about compliance with the norm. If the process is malfunctioning, the organism can’t live its normal life.

The first thing to note about this view is just how it is different from my own. On this view, the facts that ground norms are facts about the normal life of an organism and what is required to live that life. This is different from first-personal facts about mental aims. So the kind of teleology I define implies no similarity between
reasoning and processes with no first-personal mental ontology, like digestion for example. Whereas, on Foot’s view, the teleology of both is of the same sort.

Without giving anything like a full argument against Foot’s view, I’ll just note that my motivation for rejecting biological norms is that they fail the Appeal of Deontic Norms success condition.

Notice that the answer biological norms give to the question “Why ought you be instrumentally rational” is different from the answer the agent himself would give. Such a norm would say, “because it is constitutive of a normal human life”. But that fact might be completely irrelevant to a deliberating human agent. That agent would say something like, “Because I am committed to my ends”, or something like this. So the biological norms itself is not the norm the agent is responding to in reasoning—there is another internal mental norm that he finds appealing. On Foot’s view, the agent will find different types of aims appealing—like caring for others, or being instrumentally rational. But this means that her view still requires a theory like TF to explain the norms internal to the reasoning process. Talking about biological norms ends up being a non sequitur for describing the nature of deontic norms. In our terminology from the beginning of this chapter, biological norms are a kind of explanation of rational action—not a description of the nature of rational explanation.

2.5 Conclusions

Teleological Functionalism doesn’t tell us much about what the steps in practical reasoning are. It just lets us know how our psychology is aligned to aim at the ends it takes. In the next chapter, I will look at independent considerations for what the steps in practical reasoning look like. Then, I will give a teleological functionalist explanation of the deontic normative pressure inherent in those steps.
Chapter 3

The Steps in Practical Reasoning

3.1 Introduction

In the previous chapter, I argued that a theory of rationality needs to answer two questions. The first is about the nature of deontic normative pressure. In that chapter, I argued that this pressure was the pressure of teleological functional roles—the pressure to express the aims of goal-directed mental states.

The second question we need to answer is about the steps in practical reasoning, or, “what are the relata connected by deontic normative pressure?” This amounts to asking what mental states are linked in a chain from assessing the state of the world to executing rational actions. Additionally, a teleological functionalist account of these steps in practical reasoning will require a description of the teleological functional roles of the mental states involved. So in this chapter, I will investigate the steps in practical reasoning, what mental states are involved, and what the teleological functional roles of those mental states are such that they explain the deontic normativity of the practical reasoning process.

By the end of this chapter I hope to have clearly presented a precise formulation of the common sense psychological picture that theorists like Sen, Lewis, and Davidson think must justify modeling agents in terms of expected utility maximization. EUM is “...the very core of our common-sense theory of persons, dissected out and elegantly systematized” (Lewis 1974, p.338). A challenge that will occupy much of the remainder of this project is that when we try to get precise about what the common sense picture is, we find that there is actually a lot of disagreement. Lewis and Davidson, for example, present an account of rational norms that is different from the one expressed in this chapter. I will address it in chapter 4.

A difficulty that arises at this stage is the challenge of bridging the ‘is’-‘ought’ gap
in reasoning. This is the idea that if practical reasoning is about assessing the state of the world, and/or the state of myself, and then acting appropriately, there seems to be an intermediate step involving a judgment about what I ought to do. But as I mentioned in the previous chapter, ‘ought’ judgments are not obviously logically entailed by the content of our perceptions, nor are ‘ought’ facts obvious physical or metaphysical consequences of ‘is’ facts. How practical reasoning overcomes this challenge is a controversial issue, so the answer I give to it will never be fully ‘common sense’. In this chapter I will discuss this problem and my solution to it.

3.1.1 Thesis and Plan

In this chapter, I will outline the steps in practical reasoning in two ways—both by investigating what must happen in order for practical rationality to be possible, and through relying upon common intuitions about the folk psychological mental states we use in reasoning everyday. I will also provide teleological functionalist definitions of these mental state terms that facilitate the steps I outline. What I am aiming for is the naive picture of human psychology that justifies modeling agents using EUM. Even though, in this chapter, I am attempting to present a view that is as common sense and uncontroversial as possible, different considerations will require that I make two controversial assertions about the steps in practical reasoning.

The first is that beliefs, not desires, are the proximate justifier of rational action. That means that actions are more directly normative responses to beliefs about what I ought to do, rather than desires to act.

Secondly, I will claim that the nature of human perception reveals that the normative bridge crossing the ‘is’-'ought’ gap is the expression of a certain technical notion of desire, which I will specify.

I do not claim that the considerations in this chapter are a complete defense of the overall picture of practical reasoning it presents, but I do intend for this picture to be at least superficially plausible. The benefit of using the common sense picture as a foundation is that it should have at least a small presumption of correctness in virtue of how it matches the way we talk about action in our everyday lives. It would then only be necessary to discard this picture if it were incompatible with well-defended philosophical or scientific considerations. My ultimate goal with this project is to demonstrate that none of the considerations governing EUM, or objections to EUM, give us a reason to reject this common sense picture. So only at the end of this project will my defense of the content of this chapter be complete. The goal of this chapter, then, is just to develop and present a working hypothesis that demonstrates the possibility of a teleological functionalist account of rationality.
Additionally, I want to note that this chapter will leave open important questions about compliance with EUM. Specifically, it will not describe how the components of the formal model map onto the psychological picture I present. This will happen in the next chapter.

3.2 Mental States and Deontic Normative Bridges

We can describe a practical reasoning process in terms of two types of parts. There are states—things like states of the world or mental states—and there are bridges—the grounds for the deontic normative connection between states. So I might have a state $A$ and a response state required by a deontic norm, $B$. Whatever makes it the case that if you have mental state $A$ you ought to $B$ is the bridge. Another way of describing bridges is in terms of what makes that step in reasoning sufficiently universal. Universalizability is a test of the extensional adequacy of a purported normative bridge. We are looking for steps in reasoning that the right set of agents would find appealing, so we want strong universalizability. Universalizability is also related to warrant. If a step in reasoning is warranted or justified, then that is a way the step could be universalizable\(^1\).

Within the seminal works that ground this discussion, bridges seem to come in three varieties. The first is logical/analytic. If $A$ is a judgment that all men are mortal and Socrates is a man, then a candidate for $B$ would be the judgment that Socrates is mortal. The logical connection between these judgments is what makes this step in reasoning universal. Any agent that is sensitive to logical facts would appreciate that they are warranted in making this inference.

The second kind of grounds for a normative connection is physical/metaphysical. This category includes all normative bridges that are grounded in mind-independent facts about the structure of the universe. That definition is a bit fuzzy, but there are examples that can help clarify the idea. Suppose I am going to drop a ball and understand the laws of physics. I should be warranted in judging that the ball will fall. Or say, for example, that you judge something to be red. You can then infer from the fact that it is red that it has a shape. Having a shape is not analytically or logically entailed by being red—that is, the definition of red doesn’t include a clause about shape. But anything that has a color must have some shape or other. This is a metaphysical fact justifying a kind of metaphysically grounded inference. Physical

\(^1\)Another way of thinking about universalizability is as a combination of the Extensional Adequacy and Deontic Appeal success conditions from chapter 1. We are looking for deontic normative bridges that would be appealing for the right set of agents.
and metaphysical are one category because they both stem from my appreciation of facts about the way the world is independently of my own psychological dispositions. Furthermore, it isn’t entirely clear how to separate metaphysical from purely physically grounded inferences. For example, the judgment that the ball will fall if I drop it relies on induction, which itself may rely on metaphysical premises. I will leave these questions aside and just claim that there are some steps in reasoning that are warranted because they are just about recognizing that the world follows particular rules.

The third type of normative bridge is an autonomous normative bridge. Both metaphysical and logical normative bridges will require psychological dispositions in order to express themselves in reasoning. This is because the agent will have to display a kind of sensitivity to metaphysical and logical facts in order for there to be any story about how an agent acts because of those logical or metaphysical facts. But there are cases where practical reasoning requires a normative response but neither a metaphysical nor a logical fact justifies, explains, or universalizes the relevant step in reasoning. For example, the final step involved in being practically rational is actually following through and performing the action that you think you ought to perform. Logical entailments do not apply to actions themselves, but only judgments or statements about actions. So a logical bridge won’t work. Metaphysical bridges don’t seem like plausible candidates either because this final step moves from a psychological state—a judgment about what I ought to do—to an action. So it would seem odd if a mind-independent fact about the structure of the universe grounds a connection between my own judgment and my own action. On its face, it seems more plausible that a disposition that arises due to the structure of my psychology is responsible for that particular response. This step would, then, be unsupported by any fact beyond an internal psychological disposition. It is as universalizable as the requisite psychological disposition, but it might be fair to say that it is unjustified in the sense of not tracking any fact about the world.

### 3.3 Discovering the Steps in Practical Reasoning

In this section, I want to modify the common sense notion of human action using philosophical considerations about what bridges in practical reasoning could be sufficiently universal.

---

1. The author’s notes indicate that these three types of normative bridges are not exhaustive, but they are the only obvious alternatives employed in the canonical discussions of practical reasoning in the vicinity.
In order to get clearer on what the common sense view of rational action actually looks like, I’ll go through different steps that must take place in a particular exercise of practical rationality if practical rationality is about forming judgments about what you ought to do and then acting. We are answering ‘How possible’ questions about practical reasoning. As I move through, I will examine which kinds of bridges could explain the universalizability of these steps. After getting a sense of what the steps are, and why they are universal, we should have a clear enough picture of practical reasoning that we can interpret the EUM model in these terms.

### 3.3.1 An Aside: Basic Versus Complex Normative Judgments

I suggested that desires are what allow us to bridge the ‘is’-‘ought’ gap in reasoning. Views that have this property I will call *internalist*. The kinds of ‘ought’ judgments that desires immediately require us to form I will call *basic* normative judgments, or basic ‘ought’ judgments. But, on a picture like this, there may be additional kinds of ‘ought’ judgments. If this picture could be accurately represented by EUM and game theory, for example, this would imply that the logical consequences of our basic ‘ought’ judgments result in *complex* ‘ought’ judgments that comply with the demands of these theories.

This marriage of theories could be a good thing both for internalists and for EUM theorists. The reason that this could be good for internalists stems from what Mark Schroeder calls the ‘too many reasons’ and ‘too few reasons’ objections (Schroeder 2008). The core idea with these objections is just that the internalist position seems to suggest that there is normative pressure to do things that you obviously shouldn’t do (too many reasons) and there isn’t any normative pressure to do some things you obviously should do (too few reasons). For example, if you have no desire to help a starving child, then it seems to suggest that you ought not to do it. But if decision theory and game theory are models of the internalist picture, then we have the resources to clearly track the implications of the internalist picture of normative pressure in complex situations. Judgments about whether I should help a starving child need not be basic normative judgments stemming directly from desires. Rather, they might be a complex consequence, through EUM or game theoretic reasoning, of my basic ‘ought’ judgments.

While this is not a panacea for deriving all the results we might want, it can give us ways of demonstrating that the internalist picture might be more in line with complex moral intuitions than we thought. For decision theorists, partnering with

---

3(Becker 1971) is a prototype for how decision theory can be used to demonstrate that an inter-
internalism could be a good thing because internalism seems to have the best answer for how agents reason through the ‘is’-‘ought’ gap, as I will argue below.

3.3.2 Reasoning to a Particular Action

One step in the process of practical deliberation involves connecting judgments about what kinds of actions are good to judgments about the particular action you should take. This is an implication of the idea of means-ends reasoning. If you ought to perform an action because it is a means to getting what you want, then this involves connecting the fact that a particular action is a means to what you want with the judgment that you ought to perform actions that fit that description. All this is to say that a plausible schema for a segment of an appropriate practical reasoning process might look something like this:

**Deduction 1**

Premise 1.1: I ought to perform actions that bring about state of the world $\omega$.

Premise 1.2: Action $a$ brings about state of the world $\omega$.

Conclusion: I ought to perform action $a$

I will discuss different steps in practical reasoning in latter subsections, but I want to focus on the above steps for a moment. Failing to believe the above conclusion if you believe the premises is a mistake in the sense that it is a failure to believe a relevant and simple logical implication of your beliefs. We are, in a sense, already committed to the truth of the logical implications of our beliefs, which is at least pressure on us not to deny them. It also seems plausible that if those logical implications are relevant to something we are interested in, then there is positive pressure to discover them. If we are willing to accept that practical deliberation has a deductive structure, then we have reason to think that this step in practical reasoning has a logical/analytic grounds for the normative bridge. We can represent this step in practical reasoning with Figure 4.1.

Internalist approach could recommend against discrimination even though it might have been thought that internalism could not handle cases like that. (Gauthier 1986) attempts to show how internalism could recommend morality more generally.

---

4I do not have a position on whether this schema, or these normative relationships, must express themselves chronologically, from first to last, in order for the agent to be rational. It seems possible that the agent might have these thoughts out of order. Rather than tracking chronology, these relationships track what is an appropriate deontic normative response to what. More so than requiring that the agent actually think in this order, what is required is that the agent is capable of double-checking along these lines.

5The “internal” tag highlights the fact that the state underneath is part of the agent’s psychology.
Figure 3.1: Reasoning from General to Particular Actions

This is not to say that we have fully described the nature of normativity yet, or why logic is normative for reasoning at all. But we have at least reduced the question of the normativity of this component of practical deliberation to a question of the normativity of logic for reasoning.

3.3.3 From Judgments about Actions to Acting

Once an agent comes to the judgment ‘I ought to perform action $a$’, in order to be a fully rational agent he actually needs to $a$. But the connection between his judging that he ought to $a$ and $a$-ing is not a purely logical connection. This can be seen simply in the fact that $a$-ing is not a sentence—it just isn’t the right kind of object to be logically entailed by anything.

But even though $a$-ing is not logically entailed by the judgment ‘I ought to $a$’, it certainly seems normatively appropriate given that an agent has formed that judgment. Colloquially, we say that acting is facilitated by another psychological state—intending, or willing, which I will take to mean the same thing for our purposes.

Picking between our two remaining bridges, because a judgment is a psychological state, and intending or willing is a psychological process, it makes sense to think of the normative connection between judging that one ought to do something and actually doing it as autonomous, rather than as physical/metaphysical. Rational intending/willing is being psychologically disposed (in a teleological functionalist aims sense of disposed) to do what you judge you ought to do.

So in addition to the logical/analytic component of practical rationality that gets us from claims of the form ‘I ought to perform actions that bring about $\omega$’ to ‘I ought...
Judgment that I ought to perform actions that bring about $\omega$.

Figure 3.2: Reasoning from General Actions to Acting

to perform action $a'$, there is also an autonomous deontic normative connection that gets us to actually do what we judge that we ought to do. This additional step is represented in Figure 3.2.

3.3.4 From What Is to What Ought to Be

Our flow chart for practical reasoning is not complete because we have not yet described what judgments about what we ought to do are a normative response to. It makes sense to think that judgments of the form “I ought to perform actions that bring about $\omega$” are a deontic normative response to something for two reasons. The first is that judgments like that seem constrained by other considerations. For example, it would be strange if $\omega$ was a state of the world where the deliberating agent was in extreme pain for no reason. The second reason to think that these judgments are a normative response is that there are plausible options for what it is a response to. It seems that reasoning about the state of the world I should bring about is dependent upon what I think is good. Another way of putting this is that I ought to bring about states of the world because of the attractive properties they have$^7$. I’ll use the letter $G$ for whatever attractive set of properties $\omega$ has.

---

$^7$This might seem to be an endorsement of consequentialist reasoning—that value is a property of the outcomes of actions, and that I judge that I ought to perform actions because of the value that outcomes have. But I mean to define state of the world broadly here so as to possibly include properties like “Being the expression of a coherent will,” or other positions like this. In such a case, coherence would amount to endorsing a particular relationship between judgments about goodness and judgments about what I ought to do. The primary difference between a coherentist position and the one I am presenting here is that the arrows might be bidirectional. I will put off addressing objections from coherence until chapter 4.
Deduction 2
Premise 2.0: State of the world \( \omega \) is \( G \).
Lemma 2.1: I ought to perform actions that bring about state of the world \( \omega \)
Premise 2.2: Action \( a \) brings about state of the world \( \omega \)

Conclusion: I ought to perform action \( a \)

By itself, Deduction 2 may not be valid. Lemma 2.1 doesn’t obviously follow from premise 2.0 purely in virtue of the meanings of the words. But if \( \omega \)’s being \( G \) analyses to ‘I ought to perform actions that bring about \( \omega \)’, then the deduction is valid with some rephrasing\(^8\).

The problem is that no statements about what is seem, on their face, to analyze to an ‘ought’ statement. This holds whether \( G \) is some list of empirically observable properties, or even the claim “\( \omega \) is good.” This is just to reiterate one piece of the first challenge of the nature of deontic normative pressure from chapter 1—that no statements about entities we have independent empirical justification for logically entail anything about what I ought to do. But here I want to go one step further and claim that even if we allowed for an additional sui generis property that we do not have independent empirical justification for—goodness—that property still doesn’t obviously entail anything about what I ought to do as a matter of pure logic. We can see this in two different ways:

Firstly, we can compare the relationship between our candidate properties and ‘ought’ statements with archetypal cases of analyticity. For example, the relationship between goodness and the idea that I ought to do something just isn’t like the relationship between ‘bachelor’ and ‘unmarried man’. This is even more obvious with other types of properties. For example, ‘that people are suffering’ does not contain as part of its meaning ‘that I ought to prevent it’.

Secondly, one might think that a statement about goodness has a different subject than a statement about what one ought to do. Goodness, at first blush, seems like a property of states of the world—a world with no suffering is good. But a claim about what I ought to do is about me. There must be some property of me that accounts for what I ought to do, and that is a fact that is at least additional to a fact about a world with no suffering. Whatever property constitutes the fact that I ought to do

\(^8\)Another approach to constructing a valid deduction to a conclusion of the form “I ought to perform action \( a \)” is to have instrumental rationality itself as a premise, e.g. “I ought to perform actions that bring about the most \( G \) state of the world.” At first glance, this might not seem to require deriving a claim about what I ought to do from a claim about \( G \) because our premise about instrumental rationality bridges that gap for us. But if instrumental rationality was taken as its own premise judgment, then we would face the question of why all rational agents ought to believe it. It would turn out that the reason to believe that premise would still depend upon the nature of \( G \).
something, it would be actualized in me before I brought about the good state of the world—that is, before the goodness of that state obtained. This suggests that the meaning of a claim about goodness doesn’t reduce to a claim about what I ought to do because it is about a related, but different, subject matter.

None of this is meant, yet, to be a conclusive argument that logic alone can’t take an agent from perceptions of what is to judgments about what ought to be. At this point I just mean to motivate a problem. This problem is the ‘is’-‘ought’ gap. It’s the question of how we get from claims about what is the case to claims about what we ought to do when there is no obvious valid deduction between them. ‘Ought’ statements don’t seem on their face to be logically entailed by statements about the kinds of judgments we make about the way the world is. This means that if we want to capture the way in which judgments about the world are related to judgments about what we ought to do, we need to look closely at the content of our perceptions to see what they could possibly logically entail, and at other candidate deontic normative bridges.

We can start with the alternative—autonomous deontic normative bridges. Suppose that desires play the role of getting us to form judgments about what types of actions we ought to perform. This could happen either as a normative response to what we independently judge to be good, or desires could themselves explain both why we judge certain things to be good and why we judge that we ought to perform certain actions. In this second case, desires may not need to be a normative response to anything. In either case, judgments about what we ought to do are not inferred from other types of judgments in their most basic form, nor are they the result of directly perceiving that the world itself is a particular way apart from my subjective appraisal of it. On such a view, the ‘ought’ claim takes on a subjective reading—something akin to being committed, rather than being under an external obligation. On views of this type, the normative bridge to forming a judgment about what types of actions we ought to perform is autonomous—in this case facilitated by desires, rather than by intentions/the will.

The objection to thinking this connection is autonomous is that it seems to mischaracterize what many of our normative judgments are about. The idea is that when I think about what I ought to do and why I ought to do it, in many cases I think about the world itself, and not my own internal psychological states. For example, if I see a commercial about people starving in some faraway country, I may

---

9There may be cases where we infer that we ought to perform actions because of how they are related to things that we have already concluded that we ought to do. But in this case we are investigating a basic ought judgment—one not dependent upon a more fundamental ought judgment.
judge that I ought to perform actions that help alleviate that suffering. But the reason that I ought to is not because I have a desire to, but because those people are suffering. Indeed, I would still have an obligation to help even if helping wasn’t the means to satisfying any desire that I had. It seems as though if desires are what explain the normative connection between what is and what ought to be, then we must be wrong about many of the normative judgments that we regularly make.

Metaphysical normative connections are a possible way to avoid the apparent trouble with autonomous bridges for the ‘is’-‘ought’ gap. A metaphysical ground for the bridge between these judgments would imply that goodness, or some other mind independent property or properties, is linked to another mind-independent property: ‘to-be-done-ness’. We can judge that an action has to-be-done-ness because of mind-independent facts about the way it is metaphysically related to states of the world with property $G$. The benefit of an approach like this is that it can be made to fit much of our normative language. I have a reason to stop suffering because of the badness of suffering itself, not because of a fact about my own psychology.

This position is externalist because, in the right cases, being sensitive to what is objectively there in front of my face is enough for me to appreciate that I am under an obligation to do something, sans desire. I might pick up that there is to-be-done-ness by way of direct perception—directly perceiving the to-be-done-ness of an action given the state of the world. Or I could perceive to-be-done-ness indirectly—concluding that an action has this property through inferences from properties that I can directly perceive, like color or shape. For example, I might perceive patches of color that I infer are people suffering, and from facts about suffering infer that there is to-be-done-ness somehow instantiated in acts that prevent it. We will call this position Externalism ‘Minus Desires’ because it does not employ desires to bridge the ‘is’-‘ought’ gap. We can represent this position in Figure 3.3.

3.3.5 Objections to Externalism ‘Minus Desire’

I want to start this debate with the consideration that to-be-done-ness is not a physical consequence of the combination or effects of any physical properties that we can observe or measure. This is just to repeat a consideration that has come up repeatedly throughout the first two chapters. That I ‘ought’ to do something is a fact in an entirely different category than facts about how molecules are arranged and the ways they push and pull on one another. So to-be-done-ness as a property of states of the world seems to be irreducible to the physical properties given to us in the scientific picture of reality.

Teleological Functionalism addresses this problem by making to-be-done-ness a
property of a first-personal point of view. It is just my subjective commitment to an end. But a metaphysical bridge between what is and what ought to be is not a psychological fact, it’s a third-personal fact about the world itself.

So, it would seem, if to-be-done-ness is not reducible to other kinds of physical facts, then it must be its own sui generis property of the external world. One that, it so happens, we do not have the scientific apparatus to measure or detect. In The Quest for Reality, Barry Stroud argues that in the context of debates like this one, we cannot just assume that the world does not contain these additional properties: “Keeping one’s understanding of the scrutinized beliefs free of any prior conception of what is supposed to be available in reality gives the metaphysical project a chance of revealing something about the relation between those beliefs and reality, if it can be carried out. But it also makes it possible to resist the pressure towards reductionism or unmasking that any such prior conception carries with it” (Stroud 211).

I’m not entirely sure that we should reject a predisposition towards reductive theories. Firstly, there are considerations like those I raised in chapter 1 regarding the Empirical Support success condition. And secondly, if we can explain all the phenomena we observe just as well using a simpler picture of reality, then why not? Occam’s Razor seems to suggest we should go for the simpler theory. If one could present an internalist picture that relies exclusively upon entities we have antecedent reason to believe in, and that recommends the kinds of behaviors we would want to endorse, then it is fair to say that it should have a presumption of correctness over a theory that requires external ‘ought’ facts that we cannot detect or measure.

But, obviously, philosophers like Stroud do not accept arguments like this one, even if I happen to find them plausible. The scope of this project requires that I pass
by yet another debate that deserves an independent treatment. I will not attempt to argue conclusively that one must reject the physical/metaphysical deontic normative bridges that appear in externalism minus desire. But I will note an additional consideration that makes it seem implausible to me and then move on assuming internalism.

Assume that to-be-done-ness is an additional property in the world around us. The problem arises that we cannot perceive it (in the bare physical sensation sense of perceive), or infer that it is there through reasoning from what we can perceive. If this property were there, the fact that we have no access to it would imply that it can’t be involved in our practical reasoning process.

We can break down the explanation for why we can’t perceive to-be-done-ness into the case of direct and indirect perception. Direct perception would be to pick up to-be-done-ness using my sense organs and without any inference involved on my part. Core examples of direct perception are things like shapes, colors (or wavelengths of light), and temperature. This is traditionally contrasted with indirect perception—cases where I employ some form of inference to determine what is in front of me. For example, one might think that I do not directly pick up that there is a skyline before me. Rather, I infer that there is a skyline on the basis of my engagement with my surroundings and my background knowledge about what buildings are like, what constitutes a skyline, and how unlikely it is I would be deceived\textsuperscript{10}.

Direct perception isn’t a good candidate for explaining our access to ‘ought’ facts because everything that my sense organs can pick up directly is measurable by science, and we already acknowledged that to-be-done-ness is not measurable by science. So if to-be-done-ness is not an internal mental property, then I can’t directly perceive it.

The other option is that I indirectly perceive to-be-done-ness. So I pick it up on the basis of inferences from things I can directly perceive. But as I mentioned above, there is no logical connection between facts about what is and facts about what ought to be, so these inferences cannot be logically grounded. Additionally, these inferences are not simple inferences grounded in normal physical facts. My judgment that I see a skyline seems to bear a much more straightforward relationship to properties that my eyes directly pick up than my judgment that murder is wrong.

But are these inferences grounded by metaphysical truths? If they are grounded by metaphysical truths they are truths that I do not have conscious access to. The idea that murder is wrong, for example, is not a conceptual requirement in the way that ‘squares are not round’ or ‘red things must have a shape’ are. Saying that murder isn’t wrong may be false, but it isn’t complete nonsense like saying that

\textsuperscript{10}See, for example, (Fodor and Pylyshyn 1981).
you saw a round square. Whatever the purported metaphysical connection between natural detectable facts and indetectable to-be-done-ness, it is neither empirically nor conceptually accessible.

This creates an insurmountable problem for the idea that a metaphysical bridge like this could explain the universalizability of this step in reasoning. The step isn’t universalizable simply in virtue of the fact that reasoning agents wouldn’t have access to it even if it were true.

Kieran Setiya provides an alternative picture of how agents might reason through the ‘is’-‘ought’ gap even without having direct access to external ought facts or an understanding of how or why they are connected to detectable natural facts. He says that as long as external ‘ought’ facts consistently accompany natural facts we can detect, then that could be enough for an agent to be a reliable indirect detector of ‘ought’ facts. In a case like this, the justification for ‘ought’ claims would be the natural properties that the agent can perceive—things like suffering, or the means to an end, or whatever natural fact to-be-done-ness accompanies.

One issue with this view is that it makes deontic normatively strangely opaque from the perspective of the reasoning agent. He ought to do something because of a property he cannot detect, and, because he cannot detect it, he cannot demonstrate his own reliability in detecting it, even if he is reliable. Imagine a reliable detector of to-be-done-ness started to have doubts about his own reliability. There is nothing he could do to reassure himself that he was on the right track.

A related worry is that this position would fail a ‘So what?’ test for agents who should be in the extension of irrationality. A step beyond doubting my deontic normative conviction is saying ‘So what?’ to the purported deontic normative requirement. If an agent responded ‘So What?’ to the charge that he was allowing unnecessary suffering, there is nothing else we could say to him on this view. Even if it existed, any to-be-done-ness would be undetectable and impossible to present to this agent.

I do not intend to give a complete argument against Setiya’s view. I just want to note this contrast, both in the way in which the ‘is’-‘ought’ gap is bridged, and in the level of accessibility of deontic normative facts I claim is appropriate for a theory of reasoning.

3.3.6 Two Objections to Internalism

A standard objection to internalism is that it doesn’t seem to capture the way people actually ought to behave. For example, a sociopath may have no desire to help others but he still ought to. I believe that the most promising approach to addressing ques-
tions like this is through applications of game theory. If internalism is representable using these mathematical models then I believe that results in the application of those models can shed light on what the sociopath is missing out on in his inability to care about and assist others. But, for now, we are still in the process of justifying something like an internalist interpretation of decision theoretic models, so I want to bring up two objections to this position that are appropriate to our proximate goals\textsuperscript{11}.

The first objection raises two examples where agents appear to reason appropriately without employing a desire. The second case objects that treating desires as a source of normativity misrepresents the content of practical reasoning. Addressing the first objection will help give us a clearer sense of the accessibility of normative facts required for reasoning and how the different steps in practical reasoning work together. The second objection will spawn Externalism Plus Desire—a kind of hybrid position.

**Objection 1: Bob, Lenny and the Satisfaction Intuition**

I want to bring up two examples where it seems as though the agent is doing something intuitively rational, and yet his behavior does not easily fit into an internalist framework. We will call our sense that the agent did something right in these cases our *satisfaction intuition*—the intuition that the agent in this case satisfied some requirement of rationality\textsuperscript{12}.

1. After some course of experiences Bob comes to believe in an all-powerful most-perfect creator God that has a mind that is far superior to our own and that He has a plan for us that is beyond our comprehension. Bob believes that part of God’s plan is that Bob join the seminary even though Bob had no antecedent desire to do that or anything to which joining the seminary would contribute as a means. Bob decides to join the seminary.

2. George’s friend Lenny is not very bright. Along the course of a long walk Lenny catches a rabbit to keep as a pet, much to George’s dismay. In order to convince Lenny to let the rabbit go, George says, “Come on Lenny. I know

\textsuperscript{11}John Broome has a technical objection to giving an internalist interpretation of decision theory that it cannot provide for substantive constraints on preferences in light of the possibility of redifferentiating alternatives. I will address this objection, in part, by presenting a new interpretation of the nonpractical preferences that arise when alternatives are redifferentiated. For now, I want to focus on internalism divorced from questions about its formal representation.

\textsuperscript{12}This is similar to an approach Niko Kolodny takes in (Kolodny 2007) albeit framed very differently.
you better than you know yourself and I know that you don’t even want that rabbit. You want the rabbits we’re going to get later. So put that one down.” Lenny thinks about it a bit—he knows that George is very smart and rarely wrong. So if George says he doesn’t want this rabbit then he probably doesn’t want this rabbit, so he reluctantly lets the rabbit go.

In both cases it seems very natural to say that the appreciation of an external fact made a certain normative response appropriate. In the Bob case, it is a fact about God, and in the Lenny case it is a fact about what George says. In neither case is there a desire in the agent bridging the ‘is’-‘ought’ gap. And yet, something seems right about the way Bob and Lenny responded with judgments about what they ought to do from their judgments of what is the case. The question is, can this be compatible with an internalist picture? Williams is sensitive to this very issue: “Does believing that a particular consideration is a reason to act in a particular way provide, or indeed constitute, a motivation to act?... The claim is in fact so plausible, that this agent, with this belief, appears to be one about whom, now, an internal reason statement could truly be made... A man who does believe that considerations of family honor constitute reasons for action is a man with a certain disposition to action, and also dispositions of approval, sentiment, emotional reaction, and so forth” (107).

Let’s focus on Bob’s case first. The straightforward interpretation of the situation would be that Bob formed a belief about God and that that belief rationally impelled Bob to form a desire to do what he believed God wanted him to do. This is not, on its face, a case of Bob finding the means to the satisfaction of desires that he already had. However, a committed internalist might object: what actually happened was that Bob had an antecedent desire for something for which joining the seminary was a means. Bob’s newfound belief in God was just what illuminated that means-ends connection for Bob. For example, Bob might have wanted to live a good life, and once he realized that God wanted him to join the seminary, he determined that this course of action would be the best path to the good life he already wanted.

There are two reasons why this internalist description of Bob is doubtful. Firstly, given the spectacular belief that Bob has just formed, it doesn’t really need to be the case that he had any antecedent desires that could serve as appropriate candidates—perhaps Bob was a listless character who didn’t care about much at all. But once he discovered a being so miraculous and powerful and beyond his comprehension he decided that he probably ought to do what he was told. The second reason the internalist description is doubtful is that if one is still inclined to assert that Bob had an appropriate candidate antecedent desire somewhere, then given the case as just described it seems likely that Bob didn’t know what that desire was. If Bob
didn’t know what that desire was then this internalist description fails to explain how Bob acted *through reasoning*. If Bob doesn’t know that his desire to join the seminary is a means to satisfying an antecedent desire then he wouldn’t be capable of reasoning to that conclusion—this is true whether he knew he had the antecedent desire and he just didn’t realize the connection, or if the internalist attempts to assert the existence of some sort of unconscious desire. In the first case the lack of the recognized connection means a failure of the reasoning requirement in this particular case. In the second case, an unconscious desire can never satisfy the reasoning requirement. Another way of putting this is that for the internalist, the desire answers the ‘So What?’ question regarding why Bob should join the seminary. But Bob doesn’t know that he has the desire, so he can’t cite it as his answer to the ‘So What?’ question. If we have the intuition that Bob did something right here, the internalist picture struggles to explain why.

In the Lenny case it seems as though Lenny was acting rationally because he took George to be better evidence of his desires than his own introspection. This seems perfectly understandable and a case of good reasoning *given* Lenny’s limited understanding of himself and the nature of desire. Even if Lenny is an amateur internalist, it is his belief in a fact about his own desires, not his actual desires, that motivates his seemingly rational action. Once again it seems like a belief, not a desire, is the grounds for a particular bridge from judgments about what is to judgments about what ought to be.

It is worth noting at this point that Williams says some things that might seem to conflict with the picture of internal normative pressure that I am presenting here. Williams says that “A member of S, D, will not give A a reason for \([a\text{-}ing]\) if either the existence of D is dependent on false belief, or A’s belief in the relevance of \([a\text{-}ing]\) to the satisfaction of D is false” (103). In such cases we might be inclined to say that an agent doesn’t actually have a reason to act, even though he thinks that he does. Suppose that God is not real, then Bob thinks he has a reason even though he does not. Similarly, Lenny does not have a reason to let the rabbit go, even though he thinks he does.

I’ll leave aside the question of the definition of the term ‘reason’. But I do want to note that a truth requirement on internal normative pressure is problematic for two reasons. The first is that if this demand is taken seriously then we are left without a way of explaining our satisfaction intuition. Secondly, this condition makes internalist normative pressure epistemically inaccessible. An agent does not have immediate introspective access to the truth or falsehood of most of his own beliefs. If the truth or falsehood of a belief can change whether normative pressure exists for this agent, then that is a purely external condition that is changing whether
there is normative pressure or not. An agent could not be expected to be sensitive
to such a difference because, by hypothesis, he believes things to be a certain way.
So there is no way for him to know that there is no reason present in such a case.
If a standard about true belief comes into play, it is likely at a different stage in the
network of rational thought. This idea will be more clear after describing teleological
functional roles.

The internalist responses to the Lenny case and the Bob case are slightly different.
Bob has experienced a failure of doxastic rationality. If Bob found himself at the
pearly gates, he might have evidence to think that something metaphysically strange
was actually happening. But the ‘is’-‘ought’ objection would still apply. Bob would
be incapable of directly perceiving, or inferring, grounds for the demand on his
action. The nature of normative pressure would still be opaque to him. His ultimate
state of mind would have to be that he doesn’t understand why God wants him
to a, but he figures that there are some things he just isn’t meant to understand.
This model of deferring one’s own judgment in acquiescence to a spectacular event
might be warranted in extraordinary cases, but it isn’t an explanation for everyday
normative judgments. If Bob came to believe in to-be-doneness without some kind
of extraordinary perceptual evidence, he would be believing in a way that, if not
strictly incompatible with, was at least unmotivated by his best theory of the world.

The Lenny case is not so easily dismissed. But notice that even on the internalist
picture it is possible that beliefs are the proximate cause of action. If I reason from
judgments about what is good to what I ought to do, and then execute actions on
the basis of what I think I ought to do, then acting from beliefs is appropriate.
Furthermore, it seems appropriate to use the best evidence available to me to decide
what to believe. Lenny’s failure is one of appreciating what evidence is most relevant
to determining what he desires. It is only in a context where he has inadequate self-
knowledge that it is a rational response to evidence to believe George over his own
introspection.

Williams picks up on this feature of the internal/external reasons distinction when
he says that the difference between internal and external normative pressure cannot
come from examining how beliefs about what is good influence action, because both
external and internal normative pressure claims could consistently describe the same
situation. He says that it is in the ‘coming to believe’ an external reasons statement
that the peculiarity emerges and wherein we have a substantive distinction (108).

The general lesson that these examples suggest is that flawless practical reason-
ing is predicated on the idea that doxastic rationality is functioning properly. If it
is functioning properly, desires are necessary to generate judgments about what we
ought to do. Then, beliefs take over for a stage. We can reason from a judgment
that we ought to perform actions that bring about ω to the judgment that we ought to perform a particular action using beliefs and logic. From that conclusion, psychological/dispositional normative connections in intending and the will get us to act. This makes beliefs the proximate motivator for intentions. There can be mistakes at different points in the overall practical reasoning process and satisfaction intuitions in these cases come from appropriate functioning at other stages in that process.

Objection 2: Pettit and the Content of Normative Judgments

Another objection to internalism arises in that it seems to get the content of our reasoning wrong. That is, when I think about why I ought to do things, I cite features of the state of the world, not facts about my own psychological dispositions. For example, when I give to charity, I say that my reason was to alleviate suffering, not to satisfy my desires. This is to object that the internalism neglects what Philip Pettit calls desiderative structure—that we desire states of the world because of the properties that we think those states of the world have (Pettit 1991, p. 151).

That last sentence defining desiderative structure is vague between two readings. The first is that the properties that we think states of the world have normatively require desires in us for those properties. The second is that the content of my desires is specified in terms of properties. Only the first reading risks contradicting internalism, and only the second reading is necessary to explain why I cite properties as reasons for action.

The lesson from this objection for the internalist is that an agent must be allowed to attend to facts about the external world when he deliberates. He attends to what we will call the object of his desire. The object of a desire is defined in terms of properties, not fully specified states of the world. This is why the action of the play Oedipus Rex is understandable—Oedipus wants to sleep with the queen under one description but does not know that another description also holds true of her, a description under which his desires represent her as an inappropriate object of sexual interest. So a desire is about a state of the world by way of an intensional representation. When an agent cites facts about the external world as his reasons, it is still the desire that facilitated his disposition to take that fact about the world as a reason.

But why doesn’t the agent cite his own desire as the reason why he performed the action? There are two reasons for this. The first is that you do not need to know that a desire is facilitating your judgment that something is a reason in order for the desire to be operative in that judgment. A desire directs my attention to its object, not, necessarily, to the desire itself. While, on an internalist view, a claim
about goodness will always be related to some claim about desires, this means that an agent can reason competently without knowing that relationship. All the reasoner needs to do is link his judgment of something as good with his judgment that he ought to bring it about. This can happen without the agent knowing that this connection was facilitated by his desires. That doesn’t mean that an agent couldn’t reason by way of explicit reference to his desires. It just means he doesn’t have to.

However, there are two lingering problems for the internalist. The first is that he does need to pay a price by admitting that at least some of the conceptual connections people make with their judgments of goodness are false. Because desires generate assessments of goodness, those assessments are subjective—that is, they are not the result of merely observing what is there to be seen. However, there may be times when such judgments feel objective. For example, I may think that my reaction to a piece of artwork is the objectively correct one—determined solely on the basis of properties of the artwork itself. On an internalist picture, these basic judgments of goodness are not objective, and they are only as universal as the desires that create them.

Even though internalists must bite the bullet and admit that our judgments about what we ought to do are subjective, that does not preclude the kind of universalizability that we see in claims about practical reasoning, or perhaps even aesthetic or moral judgment. As long as human agents happen to have similar desires, whether as a result of internal normative pressure further upstream, or purely as a matter of similarity in construction, there will be an intersubjective universalizability of ought statements. That is, everyone who we would want to call rational will feel the appeal of deontic normative considerations. Furthermore, there may be some constraints that we must follow regardless of what desires we have—specifically, those logical constraints linking judgments about the types of actions we ought to perform to judgments about particular actions. So there may be universalizable complex ‘ought’ judgments like “you ought to take the means to your ends”.

The logical implications of our basic ‘ought’ judgments come together to form an agent’s practical stance. A practical stance represents the implications of an agent’s teleological functional aims, and would include complex ‘ought’ judgments that apply to them. I will talk about practical stances in more detail later on in this chapter.

Most ought statements are probably complex combinations of simple ought judgments, where simple ought judgments are those directly caused by individual desires. Judgments about art or morality are like this. But I will set aside attempts to explain how that is possible as it would fall under the heading of applications of the kind of model I am trying to explicate.

The second problem that Pettit’s objection raises for internalism is that maybe
the reason I do not cite my desires as the reason why I do things is that desires are normative responses to something else. This is Externalism Plus Desire.

### 3.3.7 Externalism ‘Plus Desire’

Originally, when I defined the difference between Internalism and Externalism I did so on the basis of whether a desire was necessary to bridge the ‘is’-‘ought’ gap. In the previous section, I determined that a desire was necessary for this purpose. But I want to turn to another important distinction in possible views, which is whether desires are themselves a normative response to other states, and whether those states are internal or external.

A view where desires are normative responses to external facts might suggest that things like the suffering of innocent people normatively requires a desire to prevent it. Alternatively, desires might be responses to internal facts—sensations like pleasure, pain, angst, anger, or sadness might create a normative demand to have a particular kind of desire.

I’ll distinguish these views by calling the latter internalist and the former externalist ‘plus desire’.

The internalist version of this view seems remarkably plausible. It is very hard to deny that the experience of pain creates pressure to form a judgment that you ought to perform actions to avoid it in any being that experiences pain and has the capacity to desire things. Furthermore, this version of externalism allows that desires are the bridge between judgments about what is the case and judgments about what kinds of actions an agent thinks he ought to perform. So the two remaining positions on the table are possible completions of Figure 3.4.

The argument against external normative pressure in this context is subtle, and as it will turn out, it is unnecessary to arguing that EUM is an appropriate representation of practical rationality. But for completeness’s sake, I will mention a cursory thought on this issue.

All of the normative connections I’ve listed throughout this section are ways of responding to a ‘So What?’ objection. At each step in the process as described so far, we will always have an answer to that question. For example, if an agent forms the judgment that they ought to $a$ and then don’t $a$, the problem, from their own point of view, is precisely that they ought to $a$. Or, if an agent judges that they ought to perform actions that bring about $\omega$, and then fails to appreciate the fact that there is an option available to them that fits that description, the problem, from their own point of view, is that this impedes their ability to do what they’ve already judged that they ought to do.
The problem, from the agent’s own point of view, if they fail to form a judgment that they ought to bring about $\omega$ if they desire $G$ is more complex. Answering this question will require a notion of strength of desire, and an agent’s commitments regarding what tradeoffs between desires are appropriate. Essentially, the problem that will arise if an agent fails to follow these constraints is that he will fail to form a judgment that he things it is appropriate to make by his own lights\textsuperscript{13}.

The problem with an external source of normative pressure in this context is that it risks failing a ‘So What?’ objection. Take any external fact and a purported normative desiring response, and then suppose that an agent fails to develop the requisite desire. What, from the agent’s point of view, is the problem with failing to form that desire? There isn’t any—there is no commitment or disposition within the agent on the basis of which we can say that there is a problem from the agent’s point of view.

This commitment to the idea that normative pressure must be appealing from the agent’s point of view is what does the work of denying the possibility of external sources of normative pressure. Internalist positions have the benefit of citing a built-in inclination in the agent to explain why a particular step in reasoning has universal impact. Externalists do not obviously have this option. One way of putting this idea is that there is no normative ‘view from nowhere’. All normative pressure

\textsuperscript{13}What will become clear is that I am using what Wedgwood calls a formal notion of desire on which it doesn’t make sense to ask “I have a desire, but is that a reason to do this?”
is a consequence of having a certain kind of disposition and then falling short of that internal standard. The deontic appeal success condition from chapter one, in conjunction with our assumption about the inadequacy of external normative pressure, implies that there is no normative view from nowhere.

Perhaps the externalist thinks that this should not be a requirement. Why, after all, should normative constraints be appreciable from the point of view of an irrational agent? The problem with this agent is that he does not desire what is desirable, whether he takes that to be a problem or not.

I would respond that it is exactly because we need normative constraints to be appreciable from the point of view of an irrational agent that we need to take this ‘So What?’ objection seriously. In the previous chapter, I mentioned how rationality is not just a property of rational agents, but something that exerts a pull on agents even though they are less than perfectly rational. Furthermore, because rationality is something the attractiveness of which we consciously, or cognitively, apprehend through reasoning, ‘So What?’ objections are a great way to focus our attention on what an irrational agent sees as recommending rationality. If the EPD theorist simply asserts that rationality is desiring the right kind of things, then he fails two of our success conditions—rational appeal and extensional adequacy.

The internalist variant here doesn’t have that same problem. However you want to phrase the exact relationship, pain comes along with appreciating a reason to alleviate it. Aversion to pain is just part of the experience. The problem, from the point of view of agent who does not meet an internal normative requirement to avoid pain is that he won’t take actions that could alleviate his pain—a pain that he already endorses his own aversion to.

This doesn’t imply that all desires have to be a normative response to something else. It just means that if desires are a response to something else, that something else is internal. Maybe some desires just pop up and provide their own reasons for action. For example, I might find that I want to go for a walk for no further reason than because I desire it, and that might be perfectly rational.

I’m mildly skeptical as to whether desires ever provide their own unsupported reasons in this way. Usually, it seems that we can always ask people for justifications for their desires. If I wanted to go for a walk and I am a self-aware person, usually I can justify my desire by saying it would be relaxing, or it would make me happy, etc. In any event, none of my primary aims in this project depend upon whether all desires have to be justified by another internal state. We can settle on the representation in Figure 3.6 as our benchmark internalist position for the remainder of this project, and feel free to think of the first box and bridge as optional.
3.3.8 Conclusion: The Implications for EUM

The next chapter will address the relationship between this psychological picture and EUM models. But I want to make a quick point on this subject with respect to internalism versus EPD.

While the position in Figure 3.6 is the most appealing, because of the way EUM relates to this process, it can also represent EPD. A lot more needs to be said before the relationship between EUM and practical reasoning can be made entirely clear. For example, we do not yet have a notion of when an agent makes tradeoffs between different options. But for now, notice that if preference orderings represent patterns in the judgments that we make, then the demand to maximize expected utility can express what happens downstream from the expression of desires (assuming that all this happens in a way that is compatible with EUM). This means that the normative process from either internal or external facts to desires is not represented in the structure of the model itself.

The fact that desires might be a normative response to other things is not a problem for EUM. The connection between desires and internal or external motivators can be represented in applications of EUM. As a toy example, imagine that money is the only thing that is desirable. Then only utility functions increasing in money would be admissible as possible rational utility functions. The same goes whatever the set of ultimate motivating factors. We can create hedonistic applica-
tions of the utility function that could capture this feature for the internalist position we mentioned, or a more altruistic version for EPD. Either way, EUM functions as a representation, and we are free to interpret utilities in terms of desires and preferences in each case—there are just restrictions on rational preference orderings in addition to those expressed in the axioms of EUM, which are themselves constraints on reasoning downstream from this particular step.

3.4 Teleological Functionalism and Normative Bridges

Our discussion of internalism gave us a set of roles that had to be played by some set of mental states. I will not attempt an exhaustive description of the roles of belief, desire, and intention here. But I want to articulate some of the features that these roles must have in order to facilitate internalist practical reasoning.

In addition to talking about certain kinds of mental states, I mentioned the operation of three types of normative bridges. After making some points about the teleological functionalist definition of beliefs, desires, and intentions, I will argue that the normative force of the bridges reduces to the expression of teleological functional roles.

3.4.1 Beliefs

Belief: a \textit{belief} that $G$

1. is for the purpose of representing the fact that some possible world, or the agent himself, is $G$ for access by other cognitive mental processes.

So, primarily, we identify beliefs as whatever it is that is supposed to play this role in our cognitive processes—even if, at times, it fails to successfully execute that role. In addition to the expression of its role, a belief has apparent features that are the result of its instantiation in us. For example, they have limited information storage capacity.

The definition explains why beliefs have the direction of fit they have, and, at least in broad outline, why our beliefs ought to be sensitive to evidence and what we perceive. Furthermore, as we saw in the internalist story above, beliefs play the role of facilitating deliberation by way of representing and relating certain types of content. Beliefs may repackage, rephrase, or reword information in order to interface with other mental states through reasoning. We saw an example of this in the deduction from a judgment about the general kinds of actions that an agent ought to perform.
to a judgment that he ought to perform action \( a \). The judgment that he ought to perform action \( a \) interfaces directly with intentions, as we will see in a moment.

Like desires, beliefs need to be susceptible to a method of choosing between them. This is because if they all aim at truth, and two of them contradict, then I know that one of them cannot be true. And like desires, this is made possible by certain features of the instantiation of beliefs.

Beliefs are not inherently probabilistic in their instantiation. They do, however, have a vague notion of likelihood built in. Beliefs are graded. So more precisely, beliefs have the following TF definition:

**Belief**\( G \): a *graded belief* that \( G \)

1. is for the purpose of representing how likely it is that some possible world, or the agent himself, is \( G \) for access by other cognitive mental processes.

This built-in notion of likelihood is something independent of your degrees of belief that your degrees of belief are trying to track. It is something about your relationship between your evidence and the world you are in, or the way the world is and the possible ways it could be. I won’t attempt an account of probability here. I will just note that an account of probability independent of graded belief is necessary for this position to work, so that graded beliefs have something at which to aim\(^{14}\).

This notion of likelihood is essential to the process of determining how to update your beliefs given new information, or when you determine that your beliefs contradict each other. Along with side-stepping the question of the nature of probability, I will not attempt a full defense of probabilism, the thesis that human beliefs ought to approximate a probabilistically coherent representation of the world. I just want to note here that this idea could, in principle, be defended as an expression of plausible teleological functional roles.

### 3.4.2 Desires

**Desires** A desire for \( G \)

1. is for the purposes of forming a judgment (belief) with the content that “\( G \) is a good thing to do” in the sense of “I ought to perform actions that bring about \( G \)”, where \( G \) is a property of a state of the world.

\(^{14}\)That likelihood is in this sense prior to my degrees of belief also seems appropriate from the standpoint of the content of my reasoning process. I choose actions because I think that they are effective in some sense independently of my degrees of belief. That I have a high degree of belief is not my most fundamental motivator of this kind, rather, it is what my degree of belief implies about the world and my actions in it.
2. is for the purpose of responding to internal motivating states, such as sensations and emotions.

So, by definition, a desire is a kind of disposition to form a commitment to act. In addition to this definition, desires have an instantiation in us. We talked about one of the apparent features of that instantiation when discussing Pettit’s objection to internalism. Specifically, desires themselves do not seem to be what we attend to in reasoning. Rather, desires make a state of the world seem like a good thing to bring about.

There are many different ways people use the term desire. I will relate the definition above to some of the other uses in a moment. The definition I have given here is related to what Ralph Wedgwood calls the formal notion of desire. As I will use the term, a formal desire is a disposition to judge that something is good to choose. In my terminology, to judge it as something you “ought to bring about”. The benefit of this kind of approach is that it always passes the ‘So What?’ test:

Suppose that someone raises the normative question with respect to this notion of what is a “good thing to do”: “Why should I always choose options that are good things to do? Why shouldn’t I sometimes choose options that are not good things to do?”. When the question is understood in this way, it is equivalent to the question: “Why shouldn’t I sometimes choose options that it is not correct for me to choose?” But as I explained in the first section, to say that it is “not correct” to choose an option is just to say that one “shouldn’t” choose it. So asking this question is also equivalent to asking: “Why shouldn’t I sometimes choose options that I shouldn’t choose?” But that question will hardly perplex any sane and intelligent person! (Wedgwood 2003).

Desires have the direction of fit they do because of their impact downstream in reasoning. Moving through desires and intentions, the end result of a desire is an action leading to its satisfaction.

Tracking the role of desires is a bit complicated, but it will become clearer as I move through this chapter and the next. But, in outline, once I have a set of formal desires, those desires go through a process that I will call ‘endorsement’. Endorsement is when the teleological functional role of a formal desire is fulfilled, and an ‘ought’ judgment with the appropriate content is formed. Different formal desires may be incompatible with one another. One might be for ice cream and another might be for a healthy life. But the content of the ‘ought’ judgments they spawn will reflect the strength of each formal desire in such a way that there is a logical resolution to what one ought to do all things considered.

67
In everyday language people might use the term ‘desire’ without implying any disposition to form a commitment. Cases of desires that we recognize are a result of rage might be like this. People also use the term ‘desire’ to refer to all-things-considered desires, which are the desires that we actually intend to satisfy. I’m using the formal notion of desire here because when it belongs to a natural teleological functionalist account that illuminates the steps in practical reasoning relevant to expected utility maximization. I am committed to the idea that any true theory employing the other definitions of desire will be a kind of translation of the theory I present here.

There are two reasons I treat desires as dispositions to form beliefs about what one ought to do. The first is that we need beliefs as an intermediary for influencing intentions because of cases like Lenny. Lenny formed a belief about the desires he had and acted correctly on the basis of that belief.

Lenny formed a belief about the desires he had on the basis of George’s advice. With respect to what Lenny then took his desires to be (that he wanted the rabbits he would get later) putting down the rabbit he had at the moment makes sense. Lenny’s mistake wasn’t in the action he chose given what he thought he wanted, rather, he is wrong about the desires he actually has. To do justice to our satisfaction intuition about Lenny, we had to think of beliefs as the proximate justifiers of action. Whether we behave fully rationally or not, we act on the basis of what we believe we ought to do, not just from our desires themselves. This is represented in Figure 4.2. So this means that beliefs, not desires, have the TF role of directly bringing about intentions to act. If desires are dispositions to form beliefs, then we can make sense of this picture.

The second reason to think of desires as dispositions, most directly, to form beliefs is the way in which desires must interact with beliefs in order to facilitate tradeoffs. Facilitating tradeoff behavior requires an interaction of content from the desire and content from judgments about what is likely or possible. There are three ways this kind of interaction could occur.

1. Desires could themselves neither contain content nor dispose the creation of a judgment with content.

2. Desires could be a disposition to form a judgment with content.

3. Desires could themselves contain content just like a belief.

The problem with option (1) is that the interaction between desires and the content of beliefs would not be a deliberative one. In such a case, the tendency of a desire to express itself in action could be the result of the interaction of beliefs and desires, but
it could not be a result of reasoning through the relationship between the different ways I desire things and what my beliefs say about the way the world is. I would, in a sense, just find myself inclined to perform particular actions similarly to how I sometimes just find myself with unendorsed desires. But the process of developing an all things considered intention is something we can reason through—we have some conscious role in determining what desire we choose to see fulfilled and we do so on the basis of certain types of judgments about the objects of desire. For this reason, it is plausible to think that desires are in some way generating judgments with content.

The reason that I am hesitant to endorse position (3) is because of cases like weakness of will. We often have desires without a corresponding judgment that satisfying the desire would be good, or is what we ought to do. There is an intuitive sense in which a desire like this recommends forming a judgment, but is not itself sufficient for the existence of such a judgment. For this reason, I adopt position (2), that a desire is a pro tanto, defeasible, normative disposition to form a particular judgment.

Notice that this picture can still avoid a “So What?” objection—questions like, “I know I desire this, but does that give me a reason to do it?” If we think of the disposition inherent in desires in TF terms, then a ‘So what?’ objection can’t arise. This is because the disposition in the desire is just having a TF aim—or finding something appealing from the first person point of view. So having a formal desire to X in this disposition-to-believe-something sense is to already endorse the idea that you are inclined to judge that you ought to X. But even among desires we choose to endorse, unless you’re very lucky, or unambitious, you can’t take actions to satisfy all of your desires—you have, at times, to choose between them. It also seems like this decision is something I can reason to. When I introspect on my desires, or reflect upon the things that I want, certain facts about how much I want an outcome, or how important it seems to me become apparent. I contend that, similar to graded beliefs, strengths of desire are part of a desire’s teleological functional role. In order to interface with the content of graded beliefs through conscious practical deliberation, graded desires express themselves in

---

15 This relationship is similar to a relationship that arises between the beliefs ‘I have overwhelming evidence that A’ and a belief in A itself. I can believe that I have overwhelming evidence that A, and that implies a kind of disposition or inclination to believe A, but it is itself a different mental state than believing A. Having a formal desire is similar in the sense of being in a state where you endorse the idea that you have reasons to form a belief, but it is not the belief itself.

16 The instantiation of this role happens to make strengths of desire all finite relative to one another such that some change in the probability of satisfying one is just as desirable as some change in probability of satisfying another. I will discuss this more in depth when I argue for the appropriateness of continuity as a constraint on mathematical representations of rational choice.
qualified content in ‘ought’ judgments.

Desire$_G$: a graded desire for $G$

1. is for the purposes of forming a judgment (belief) with the content that “$G$ is a good (to degree $U$) thing to do” in the sense of “I ought (to degree $U$) to perform actions that bring about $G$”, where $G$ is a property of a state of the world$^{17}$.

2. is for the purpose of responding to internal motivating states, such as sensations and emotions.

There are multiple things that ‘to degree $U$’ could mean, and, therefore, multiple ways in which a graded desire would interact with a degree of belief regarding an action. EUM and Risk-Weighted Expected Utility Theory both imply an alternative content here. I will talk about the EUM content of these judgments in chapter 3, and the REU alternative in chapter 5.

What is important for our purposes here is that ‘to degree $U$’ corresponds to a sense of how important an outcome is to the agent. It arises as a result of the strength of the graded desire that spawned it and it governs, among other things, what costs I am willing to pay and what risks I am willing to take. For example, you shouldn’t be willing to give up something that is more important for something less important. Given the logical/analytic normative bridge at this stage of practical reasoning, this means that the content of ‘to degree $U$’ must ultimately analyze to something such that there is a valid deduction from the judgments recommended by my endorsed desires, and judgments about action efficacy, to judgments about what actions I ought to take that remains faithful to these requirements. For EUM, ‘to a degree $U$’ will analyze to a statement about the contributing value of a $p$ probability of an outcome to an action. I will discuss this in detail in the next chapter. But, as I mentioned, even if you assume a TF interpretation Risk-Weighted Expected Utility Theory will have a different answer to this question.

But why think that we have graded desires at all? An alternative view, ordinalism, takes it that we have only ordinal psychological states. So I know that I desire $X$ over $Y$, for example, but that I desire $X$ “twice as much” as $Y$ has no meaning beyond being an artifact of my risk behavior. In other words, through the Von Neumann-Morgenstern axioms I can construct a graded utility function that

\footnote{So I mean ‘ought’ in a sense that comes in degrees. What you ought to do all things considered is the thing that has the highest $U$. This doesn’t mean that you neglected a duty in passing up something that you ought to do so some degree. Rather, you passed it up because it was overridden by something else.}
represents my preferences or desires over lotteries, but that is just a consequence of having coherent and unexploitable attitudes towards risk, and not the result of anything graded in my head.

My full reply to this position comes in two parts. Firstly, notice that the above position requires that there is a kind of aversion to incoherence or exploitability that is not obviously the expression of the teleological functional role of ordinal desires, beliefs, and intentions themselves. As such, it becomes difficult to understand how to explain the magnetism, or normative pressure, associated with coherence. Coherence norms of this type fail a ‘So What?’ objection, as there is nothing about the agent’s psychology explaining their attraction to coherence itself. I will discuss this more in chapter 4, which is exclusively about coherence norms and their alternative.

Here, I would just like to note a positive reason to believe that we have graded desires. The point in favor is that many of our everyday judgments are difficult if not impossible to explain using ordinal desires alone. I’ll give three examples.

Our first case is regret. Not regret in the sense that an agent believes he made the wrong decision, but a feeling that it was very unfortunate that a particular thing had to be sacrificed. If we only have ordinal desires, then what was sacrificed was sacrificed for something preferred, and that’s all there is to it. Because there is no risk involved in this decision, there is no sense in which what was given up was ‘almost as good as’ what was acquired. So, for the ordinalist, the cases where an agent struggles to give something up for an alternative are no different than the cases where the choice is obvious. Strengths of desire can help explain this phenomenon by talking about how both options were desired to a very similar, or high, degree.

This story becomes even more plausible when we recognize that strengths of desire can be vague. What I mean by a strength of desire being vague is that a range of different ‘ought’ judgments are normatively appropriate responses to one desire. A desire will not tell you the difference between a utility of 5 and 5.00001, for example. Vagueness of strengths of desire leads to the cases where an agent is struggling to figure out what he wants more. This happens only when the options are close in value, not when they are valued at wildly different levels. Because strengths of desire are vague, there is a bit of flexibility in what rationality allows an agent to do. For example, you may find some slippage in your level of desire for tea with x or x + 1 grains of sugar in it (Chang 2002). This is not a problem for rationality, but it does mean that utility functions are overly precise approximations of psychological preferences, multiple similar utility functions being appropriate representations of a person’s preferences. I will argue these points in more detail in the next chapter.

---

18This is the real explanation for slight sweetenings puzzles as seen in (Chang 2002). Notice that parity occurs in cases where it is difficult to tell what you want more. There are no cases of obvious
It will turn out that an interesting implication of vagueness is that an agent only needs to cohere to one utility function if there is an actual risk that she will be Dutch Booked.

Our second example is the case of getting a good deal. Sometimes we feel like we would have sacrificed much more than it seems we will have to to get what we want, and sometimes it seems we will have to pay all we are willing to pay. Determining this is not likely achieved by running through the list of all the things that I would trade for what I am about to get, and all the differences in sets of goods that represent the difference between what I’m paying and what I’m receiving. Rather, I have a fairly direct sense by comparing two alternatives that one is worth much more than the other, and that I am getting surplus through the trade. Once again, graded desires are the most natural way to capture this example.

Our final piece of evidence is linguistic. People straightforwardly verbally express judgments that correspond to their strengths of desire. People say things like, ‘X is much more important to me than Y’. While this statement might have implications about what rational choices over lotteries they might make, this is not in itself a statement about their lottery preferences. Someone might not know anything about decision theory and still say that they would much rather go to a baseball game than go to the dentist.

But there is something correct about what the ordinalist says about the claim that there is nothing real about X being ‘twice as good as Y’ inside the agent’s head. While it is true that desires are graded, X is only twice as good as Y relative to a scale of measuring strengths of desire. There could be multiple types of scales, and ‘twice as good as’ smuggles in mathematical relationships that are specific to the scale and not part of strengths of desire themselves. Temperature works in the same way. Today is only ‘twice as hot’ as yesterday relative to a scale, so ‘twice as hot as’ is an artifactual property. But that does not mean that there is not some real sense in which a graded property underlies those measurements. In the case of the Celsius scale, ‘twice as hot as’ originally got meaning from the impact of temperature on boiling water and the volume of mercury. For EUM, it will turn out that ‘twice as good as’ gets meaning from the impact of strengths of desire on the lotteries that we are willing to accept. Strengths of desire might have other consequences—such as tending to bring about emotional responses of a particular type. Strengths of desire also might be measurable on different scales. It just so happens that the scale we will use tracks a specific kind of implication of strengths of desire and generates a scale on the basis of recommended risk behavior.

parity, just cases of unclear betterness.
The ultimate output of an agent’s graded desires is a network of interrelated ‘ought’ judgments with strengths, all based on appropriate tradeoffs around risk and costs. Not all these judgments need to be explicit—we only need to reason with the ones that are relevant to the situation at hand. A practical stance is the collection of all the qualified ‘ought’ judgments, implicit and explicit, that a rational agent would make as expressions of their endorsed desires. In the next chapter I will argue that a utility function is most directly a representation of an agent’s practical stance.

3.4.3 Intentions and the Will

Intentions: Intentions are

1. for the purpose of causing me to perform action \( a \) from a judgment that I ought to \( a \).

Intentions, like beliefs and desires, admit of a more detailed description than I will give here. But note that without intention, none of the roles we have defined thus far could explain that we are actually under subjective pressure to do what we think we ought to do. With this role included, the interaction of teleological functional roles is enough to result in practical reasoning as described in Figure 3.6.

3.4.4 Teleological Functionalism and Normative Bridges

The three types of normative bridges, logical/analytic, physical/metaphysical, and autonomous, are themselves just ways in which teleological functional roles express themselves.

Logical/analytic and physical/metaphysical grounds for normative bridges are an expression of the role of beliefs. Both logical and metaphysical facts have implications for what is true given what we believe—e.g., if I believe that this object is a square, then that implies that it is not round, and it also implies that the object is not a square. These kinds of external world relationships are relevant to the goal of beliefs in trying to represent the way the world is, but also in trying to represent the world under different relevant descriptions. This applies to the inferences we make in deductive reasoning, and, in the case of practical reasoning, provides the relevant input for intending to act.

Autonomous normative bridges are the expression of teleological functional roles that are not themselves attempting to reflect a pattern in the mind-independent
external world. This can be seen in the expression of the TF role of desires and intentions.

It should be becoming more plausible that normative pressure is nothing above and beyond the expression of teleological functional roles. No doubt these roles need to be more precisely described in order to capture all that goes into rationality, but, for now, I only want to support the idea that the interaction between goal-directed mental states generates the relations and pressures we are talking about when we talk about rationality. If normative pressure is just the interaction of mental states with the definitions above, then any agent that is up to those things (has those mental states) will feel the force of normative pressure.

One of the consequences of a view like this is that agents who do not have mental states like these will not feel any normative pressure to be rational in the way that we normally conceive it. For example, if an agent has no desires, or if their mental states aim at something other than truth. Furthermore, even compatible with the broadly defined teleological functional roles above, there are different things that rationality could amount to. I still have to specify what judgments desires incline us to form, for example. EUM and, with a few caveats, Risk-Weighted Expected Utility Theory, can be interpreted as different ways of answering this question. But this means that if a Risk-Weighted Expected Utility Maximizer were to exist, he would not feel any pressure to be EUM. Rather, being an REU maximizer is an entirely different way for mental states to aim at performing actions. From the perspective of a true REU agent, REU itself is beyond reproach. This is consistent with two recurring themes in this project. The first is that it may be possible for there to be multiple kinds of rationality. The second is that there is no normative view from nowhere. The fact that normative pressure relies upon these features of an agent’s psychological construction (in addition to the fact that desires bridge the ‘is’-‘ought’ gap) means that there are psychological preconditions for having reasons to be other than how you actually are. The REU and EUM agent will have different assessments of what counts as a failure of rationality, and there is no perspective outside of any decision theory from which to claim that an agent ought to be one way or the other.

What this means is that arguments for a particular decision theory must be, to some degree, empirical. We need to find a way of assessing what our reasoning process actually happens to be like in order to determine whether EUM or some alternative is an appropriate description of the normative pressure in us. This, I believe, is what is actually happening when people give arguments that refer to an agent’s exploitability by a Dutch Book. Exploitability by itself doesn’t generate normative pressure. Only relative to a way of seeing reasons is exploitability normatively unacceptable. Exploitability just seems very unattractive to us, and as such is a way
of arguing that we are not the kinds of beings representable by a decision theory that allows this. This does not imply that there couldn’t be an entity whose teleological functional roles led to exploitable reasoning.

Because it seems obvious that people deviate from rational behavior in practice, another way to highlight the strangeness of a decision theory is to investigate the nature of the concepts we employ in reasoning about what we ought to do. This would be to examine the kinds of judgments we make in the space between having desires and forming intentions. This is the approach I will take in chapter 5.

3.5 Formal Modeling and Resolute Choice

One of the most impactful implications of the formal notion of desire is that resolute choice is never a case of acting against one’s preference ordering at the time of action. Resolute choice is a way for an agent to bind his future self to an action plan developed at an earlier time (McClennen 1990). This is like the real world equivalent of Ulysses binding himself to the mast as he approaches the Sirens. Resolute choice is a strategy for handling situations where you expect that your future preferences are incompatible with what you now want to happen at some time in the future. So, like Ulysses, you take action to prevent your future self from derailing your plans. The difference between resolute choice and the Ulysses case is that there is no available external constraint—the rope is all in your mind, so to speak. Resolute choice is just making a resolution to stick to your current plan regardless of how you feel about it later. If you have sufficient willpower, this is a way of avoiding outcomes that seem unappealing from your present perspective.

I immediately want to rule out the thought that resolute choice is a kind of induced non-rationality or thoughtlessness on the part of the agent. Someone might think that one way to ensure that you follow through with your plans is just to avoid thinking about them after you’ve decided what to do. Then, when you are at risk of reneging at time $t_2$ you just won’t think about the situation enough for your $t_2$ preferences to affect your action.

I don’t think this is what people are after when they say that resolute choice is a rational strategy. Firstly, this seems too similar to an external constraint. In a way, it is just removing the ability to choose later on—just like Ulysses binding himself to the mast—and, as such, wouldn’t require a separate theoretical treatment. Additionally, induced non-rationality wouldn’t be a reliable strategy. Personally, I

---

On my theory it will turn out that this aversion to exploitability is not a sui generis desire people happen to have, but a consequence of having desires like ours.
cannot make myself not think about something for very long. At times it even seems like I have to think about it in order to will myself not to think about it. Ideas just pop up beyond my control, especially when they are relevant to the situation at hand. If resolute choice was just my ability to prevent myself from reevaluating my plans in the future, it wouldn’t be a general solution to sequential choice problems.

(McClennen 1990) argues for the rationality of resolute choice by demonstrating that in certain cases a resolute chooser will end up better off than a non-resolute chooser. Our leading candidate non-resolute agent is the sophisticated chooser. An agent who uses a sophisticated choice strategy is one who uses backward induction to figure out what to do. He will consider his preferences at a time and expect himself to act in accordance with those preferences at that time. Given his expectation of his own future behavior, he will make a choice now that maximizes the expected satisfaction of his present aims.

One way to put this is that the sophisticated chooser adheres to what McClennen calls Separability:

...the agents future preferences among options are shaped by considerations that abstract from, and thus are separable from, the background of earlier preferences. More specifically, future preferences are shaped by a concern for just those consequences that remain realizable at that future time” (McClennen 1990, pp. 12).

In the actual story, Ulysses’s preferences are separable. He knows that upon hearing the Sirens he will develop preferences that disregard his current interests. So in order to avoid the expected outcome of that state of affairs, he binds himself to the mast.

But a resolute Ulysses can reach Ithaca without the inconvenience of being tied to the mast. Because he does not abide by Separability, his future preferences can take into account his past resolution to sail to Ithaca. If he is able to do this, he can reach Ithaca without paying the additional cost of being tied up—a better overall outcome than reaching Ithaca after being bound. In this way, McClennen argues that an agent can rationally justify resolute choice by way of a commitment to reaching the best outcomes.

The obvious challenge to resolute choice is that it simply isn’t feasible. Once Ulysses hears the Sirens’ song, he will prefer to steer his ship towards the rocks. If rational action just being maximally effective at satisfying your preferences, then how could Ulysses fail to pursue the Sirens and remain rational if that is what he wants to do at that time? The question becomes, to what extent, and in what contexts, can a person rationally motivate themselves to be resolute. For example, (Gauthier
argues that the rationality of resolute choice depends upon a broader context of related preferences.

But Separability is ambiguous between two readings—Formal Separability and Temporal Separability.

Formal Separability is the idea that a rational agent will make the same decisions at a node of a decision tree regardless of whether that decision is embedded in a larger tree or not. I will argue that this is not so much a question of whether agents can make resolutions as it is a question of how we construct decision trees—specifically, it is the result of either treating resolutions as a change of preference ordering, or an opportunity to redifferentiate outcomes.

Temporal Separability is the idea that an agent can’t take his past interests into account when deciding what to do.

Formal Separability implies that Ulysses may be perfectly free to take into account his past resolution. Temporal Separability implies that he is not. There are four points I want to make about Separability:

1. McClennen and Gauthier are not talking about Formal Separability.

2. Formal Separability, and not Temporal Separability, is relevant to EUM given the interpretive considerations I’ve been discussing.

3. Formal Separability is perfectly compatible both with resolute choice and captures the intuition behind the feasibility objection.

For reasons 2 and 3, I endorse Formal Separability. And because McClennen’s examples are intuitive to me, I reject Temporal Separability. This means my position implies the thesis of this section: that, properly modeled, resolute choice never implies that an agent acts against the preference ordering he has at a given time.

### 3.5.1 McClennen and Gauthier

I do not intend to give anything like an in-depth description of McClennen or Gauthier’s account or justification for resolute choice. I do this mostly because the point I want to make is orthogonal to their own. But it might help bring focus to what I am talking about to contrast it with the considerations that motivate them.

As mentioned above, McClennen is interested in the benefits that can accrue to an agent who can resolve to stick with a plan. This is incompatible with the Temporal reading of Separability. Temporal Separability might be defended by someone who thinks that the quality of an outcome is independent of the context or history that leads to it. A view like this could recommend only focusing on one’s future prospects
and leaving considerations about the past out of any deliberative process. In this way, McClennen has opponents, but not in virtue of rejecting Formal Separability.

Gauthier’s objection to McClennen is motivated by a concern that resolute choice is not obviously rational in all cases. He gives the example of a boy who is grossed-out by girls standing resolute against his expected future preference change. In this case, Gauthier argues, remaining resolute doesn’t seem like a good idea (Gauthier 1997, pp. 18-19).

All of these considerations regard Temporal Separability—how future me relates to the interests of past me. This is just a different question from the question of whether we construct decision trees to incorporate resolutions in either outcomes or preference orderings.

McClennen briefly considers Formal Separability but moves on given that Formal Separability doesn’t raise any objection to his view. He says that in a theory where the notion of a consequence is defined in the ‘inclusive sense’ (a sense that takes note of the history from which the outcome derived), the normative relevance of Separability is severely restricted, and restricted in such a way that support for Separability is not an objection to his view (146).

In my framework, McClennen and Gauthier are debating about the process by which you come to have whatever formal desires you ultimately have—in what contexts my formal desires should incorporate my past commitments. I will discuss this in more detail momentarily. But in terms of our chain of practical reasoning, this means they are debating about the process that takes a person from the first box to the second (See Figure 3.6). To rehearse an earlier point, I do not think that EUM is about this process.

### 3.5.2 Formal Separability and EUM

I want to assume from the outset that there are at least some cases in which a rational agent can do things that they otherwise would not have done if they hadn’t resolved to do them. If we don’t make this assumption, resolute choice is impossible and there is nothing interesting to talk about here—an agent just maximizes the preferences they have at a given time and there is no live opponent to sophisticated choice. As I mentioned, the idea that I can resolve to do things and then act from my resolution also seems overwhelmingly plausible to me.

Within theories that are similar to expected utility theory, there are three options for how we model situations like this. Firstly, resolute choice could be modeled as a case of an agent acting against his preference ordering. Secondly, it could be modeled as a case of preference change, where a resolution gives an agent different
desires. Finally, it could be a case of redifferentiated alternatives, where a given outcome resolved has a different utility than that same outcome not as the result of a resolution.

My claim is that we have to reject the first option for two reasons. Firstly, preference orderings must represent formal desires in order for decision theories resembling EUM to provide complete rational explanations. Secondly, resolutions imply a difference in formal desires, not acting against one’s formal desires at a given time.

With respect to the first point, the long quote from Wedgwood articulates how the formal notion of desire is sufficient for passing a ‘So What?’ test. Furthermore, other notions of desire will not pass this test unless they are augmented with additional content. Suppose that a desire does not imply a degree of commitment to bringing about an outcome. This is a natural, but incompatible, meaning of the word ‘desire’. I might say, “Yes, I know I desire X, but is that really a good reason to bring it about?” So there is a gap in reasoning from desires that do not include a commitment to action. Any theory of rationality that uses a notion of desire like this will be incomplete (in the sense I mean) unless it includes some description of how to bridge this gap.

But none of the formal decision theories in the vicinity of expected utility maximization include a structure that could capture a bridge like this. If we take preference orderings to represent desires, then a formal decision theory that doesn’t
represent formal desires will need a list of caveats about when an agent should maximize their preferences and when they shouldn’t. On a view like this, that an agent’s preferences recommend something would not be a complete explanation for why he ought to do it. Rather, his preferences recommend something given that they pass an additional set of standards. But no formal decision theory I am aware of is structured this way. Generally, theorists attempt to justify decisions as expressions of the content of whatever is represented by the preference ordering itself—incorporating all rational considerations, EUM or not, into that preference ordering.

Of course, none of this is to say that there couldn’t be a decision theory that works differently. It would just need a psychological interpretation that captures the interaction of non-formal desires and whatever additional rules. But this theory would look very different from the theories available in the literature. Additionally, a theory like this wouldn’t necessarily disagree with the theory I am presenting here, it would just highlight different parts of the decision process.

So as long as we restrict ourselves to theories resembling EUM—where decisions are justified on the basis of the preference ordering itself—then preference orderings need to represent formal desires.

With respect to the second point, that resolute choice is a case of changing formal desires, not acting against the formal desires one has: notice that if, at time $t_2$, the agent wants to follow through with his plan, then this is what his formal desires recommend doing at $t_2$. If it wasn’t what his formal desires recommended, he would face a ‘So What?’ objection from himself at $t_2$, “I know I planned a resolute choice, but why does that matter to me now?” In the absence of induced non-rationality, the situation at $t_2$ will make his preferences at that time salient to him, and he will figure out what he wants, all commitments considered, to do. If it is possible that he follows through with his plan, then it must be that his formal desires changed, from in aggregate recommending one thing at $t_2$ to recommending something else as a consequence of his resolution.

On a view like this, resolute choice is modeled as an additional branch at the beginning of a decision tree. Down one branch I have the original decision problem without having made a resolution. Down the other branch I have the decision problem given that I have made a resolution. The resolution might change the utility values of the outcomes at the end of the resolution branch of the tree—counting as a redifferentiation of outcomes. Or, the tree might imply that the agent’s preferences over the same set of outcomes change from what they would have been at $t_2$ minus the resolution. Whether we have a case of preference change or a need to redifferen-

---

20I will talk in more detail about the formal representation of this theory of rationality in the next chapter.
tiate outcomes depends upon whether the agent at \( t_1 \) has a preference between an outcome as the result of a resolution versus that outcome without having made a resolution\(^{21}\).

All of this is completely orthogonal to the debate between McClennen and Gauthier, or between both of them and someone who rejects the possibility of sticking to resolutions.

### 3.5.3 Formal Separability and Feasibility

Even though Formal Separability is compatible with making resolutions, it is all we need to capture the intuition behind the feasibility objection. That objection is grounded in the worry that it is not possible for a rational agent to act against the preferences they have at a particular time. Whether that is plausible depends upon whether ‘preferences’ tracks formal desires. If preferences are about formal desires, then it is very plausible. There is a way of hearing the sentence “I did something other than what I wanted to do” as almost nonsense. If ‘want’ here means formally desire, then we are left wondering how the agent motivated himself to act against his desire. But if ‘want’ is taken in a less inclusive sense, then this is not strange at all. For example, what I want to do in the formal sense is exercise this morning, and what I want to do in this other sense is stay in bed. Even if an observer is impressed by my resolve, no one would question the psychological possibility of rationally acting against what I ‘wanted’ in this latter sense. So the objection from feasibility is only plausible if it tracks formal desires. But our theory of formal desires doesn’t restrict their capacity to be for finely individuated outcomes, nor does it preclude the possibility of preference change in the face of a resolution. So as long as we are modeling formal desires, we can have Formal Separability and resolute choice in one package. Nevertheless, resolute choice will never be a case of acting, at a given time, against the preference ordering one has at that time.

\(^{21}\)This type of preference seems ‘non-practical’. That is, How can I ever have a choice between the same outcome as the result of a resolution and not as the result of a resolution? There is a discussion in the next chapter on ‘non-practical’ preferences and the types of decision trees that reveal their utilities.
3.6 Conclusions

3.6.1 The Experience of Reasoning

One might object at this point that I do not reason by way of thinking about the execution of teleological functional roles, and that this implies that my account of reasoning is false. But none of what has been said implies that this is what reasoning is like. As mentioned with the case of desire, I can attend to the object of my desire and then feel disposed to judge that I ought to bring it about. This is what the TF account of reasoning says that this stage in reasoning is like, and that is very different from contemplating the teleological functional roles of my mental states themselves.

In some cases people act rationally even without explicit reasoning. My phone might start ringing and I reach in my pocket to answer it, but I do so without consciously taking stock of my desires and the most effective means to satisfying them.

In cases like these I would say that my mental states are implicit, rather than explicit, to follow a distinction made by (Harman 1986). That is, there are certain beliefs, desires and intentions, that I may have before my conscious mind, which are explicit. Then, there are others that operate without my conscious attention and are implicit. The process is a lot like breathing in the sense that it can operate without my conscious control, but when I attend to it, the steps in the process become deliberate and conscious. Either way, the TF account is that these mental states have aims that express themselves in practical rationality.

All this implies that the experience of reasoning is very different from the experience of contemplating teleological functional roles themselves. As we’ve seen in this chapter, revealing teleological functional roles happens by way of asking ‘how possible’ questions about reasoning. Once we have these roles defined, we have a vocabulary capable of helping us carve up and describe our reasoning process.

3.6.2 No Normative View from Nowhere

An important consequence of the view presented in the last two chapters is that there is no normative view from nowhere. By this I mean that all goodness or badness is relative to a perspective. In order to be deontic, a normative response has to be the expression of a teleological functional role, which is a psychological state. What agents aim at, and what effective aiming amounts to, are consequences of their subjective psychological construction. This idea will arise later on when I investigate the plausibility of alternative internalist theories of rationality. In each case, the idea that there is no normative view from nowhere will require that a theory of rationality

82
completely captures what is good or bad about a particular way of acting. So if a
theory returns the result that an agent is exploitable by a Dutch book, for example,
then that theory implies that there is nothing bad about exploitability from this
point of view. As such, the agent in question would see no reason to try to avoid
exploitability. This will have implications for theories that attempt to use holistic,
resolute, choice as a way of avoiding Dutch books.

While the view presented here is a kind of subjectivism, it doesn’t share many
of the characteristics of alternative subjectivist positions about deontic norms. For
example, the demand to take the means to my ends or to be a good person does not
need to be just a matter of personal preference like choosing a flavor of ice cream.
What differs between these kinds of judgments is their universalizability. Taking the
means to my ends, and hopefully moral norms down the line, are consequences of
the structure of teleological functional roles themselves. So all agents with beliefs,
desires, and intentions will be subject to these norms. That is different from having
a taste for vanilla ice cream, which is only as universal as the desire that motivates
it.

3.6.3 Next Steps

In the next chapter I will articulate how this psychological picture can provide an
interpretation of the formal structure of EUM models—both by mapping the com-
ponents of the model onto psychological states as well as by describing the content
of our basic ‘ought’ judgments.

Chapters 4 and 5 will address objections from alternative pictures of rationality.
Chapter 4

An Interpretation of Expected Utility Theory

4.1 Introduction

At this point in the project we have an internalist account of practical rationality in hand. Teleological Functionalism is a method for defining mental states in terms of the teleological functional roles that they play in getting us to aim at practical rationality. If we can interpret the components of expected utility models—preference orderings, outcomes, and the pressure to comply with the axioms—in terms of this psychological picture, we will have presented a realist interpretation of EUM. A realist interpretation of the model is an interpretation such that the components of the model are about real world phenomena, and not just artifacts of a method of measurement.

The term ‘realist’ is itself a bit misleading. Really, interpretations of formal models occupy a gradient from fully realist to fully constructivist, given what parts of the model are meant to correspond to real world phenomena and what parts are purely artifactual. I will present a view where an agent could be represented equally well by any linear transformation of a utility function that represents him. So the fact that a particular application represents his utility for apples as 5, for example, is artifactual. But utilities, the distances between them, outcomes, and the axioms will all represent psychological facts.

The benefit of a realist interpretation of EUM is that it grounds EUM as a

\[ v(\phi) = 0. \]

This could be used to lend an additional dimension of realism to the utility function, but I will not discuss that here.
presentation of rational explanations. This is just to live up to Sen’s demand that motivates this entire project. Behavior is not inconsistent with other behavior, but with the reasons for it. If EUM can be interpreted in terms of the reasons for action, then it presents an explanation for the rationality of compliance.

But, as at every stage, there is a problem. Several authors make arguments that amount to the claim that internalist accounts of deontic normative pressure are incompatible with EUM because of the tension between Characterizational and Normative Representationalism². Characterizational Representationalism is the idea that “Notions like belief, desire, and utility do not need a firmer psychological grounding because the representation theorem [EUM] provides the necessary characterization of these states.” Normative Representationalism is the idea that EUM displays justifications for rational action, or rational explanations (Meacham and Weisberg 2011, p. 2). The problem is that cognitive science and neuroscience are building a body of evidence that seems to imply that our mental states are nothing like what folk psychological terms would suggest: “CR is implausible because it is in tension with psychology, both folk and empirical. NR is problematic because it implicitly depends on psychological claims similar to those made by CR” (3)

The basic idea is that providing a realist interpretation of EUM requires that EUM accurately represents the way our mental states are—otherwise EUM couldn’t depict the reasons we actually have. But there is evidence that purports to show that people do not psychologically or neurobiologically resemble EUM.

This wouldn’t be a problem if we weren’t internalists. If to-be-done-ness, or ‘ought facts, were mind-independent external facts, then the fact that our psychologies did not match EUM wouldn’t be a problem for deontic norms so long as these mind-independent facts matched EUM. But I discussed reasons for rejecting this view over the past two chapters.

So our questions is, how can EUM be an accurate representation of human psychology in light of this experimental evidence?

4.1.1 Thesis and the Plan

There are three tasks to complete in this chapter. The first is to present an interpretation of the formal components of expected utility theory in terms of the psychological picture presented in the previous two chapters. The second is to describe the content of our basic ‘ought’ judgments (the ones that our desires directly recommend that we form) and how those judgments logically entail compliance with EUM. The third is to examine the experimental evidence against EUM as an accurate characterization

²Beyond (Meacham and Weisberg 2011), which I will discuss at length, see (Bermudez 2009).
of human psychology and evaluate it in light of teleological functionalism. The next four sections will meet each of these goals in order, with one section dedicated to the content of basic ‘ought’ judgments and another section for compliance with the axioms. I will end with a final section on conclusions and challenges we still have to meet.

The following section is more of an explanation than an argument for a particular thesis. But sections 3 and 4 argue for the following:

- The content of our basic ‘ought’ judgments is an answer to the question “To me (subjectively), for me (as grounds for my own action), what is the value contribution from efficacy at bringing about $\omega_1$ relative to efficacy at bringing about other outcomes?”. A formal translation of the answer to this question implies all the axioms of EUM.

- Teleological Functionalist provides the relevant sense in which a psychological account needs to be an accurate characterization of mental states in order for EUM to be a good representation. The experimental evidence exclusively examines causal and third-personal information, and, as such, misses the perspective of deontic normative characterization.

4.2 Interpreting the Formal Model

4.2.1 Preference Orderings and Utility

Graded desires motivate judgments that we are under a certain degree of pressure to perform actions that bring about attractive states of the world. They might also be negative—inclining us to judge that we’d really rather something didn’t happen. While desires are directed towards states of the world in isolation, and so too are the basic ‘ought’ judgments that we form in response to a desire, preferences are essentially comparative. Preferences, and the formal preference orderings that EUM contains, regard the logical consequences of the basic ‘ought’ judgments that desires incline us to make. So, for example, if I am inclined to form a judgment that I want to bring about $\omega_1$ to degree $X$, and $\omega_2$ to degree $Y$, where $Y > X$, then I will conclude that I prefer $\omega_2$ to $\omega_1$.

Determining the amount of pressure on me to do something—the strength of my reason to do it—is like assessing something’s worth. Whether something is worth a lot or a little is not exclusively understandable in terms of what I will trade it for. Rather, I can understand the idea that I want something a lot or a little from the strength of my desire alone.
But this is not to say that wanting something twice as much as something else is something I can know exclusively by thinking about my strengths of desire. Wanting something twice as much assumes the existence of a scale that gives meaning to the numerical comparison. One of the effects of a strength of desire has to do with the kinds of risks I am willing to accept. Features of these risks allow for the creation of a scale that can give meaning to the idea that I desire \( X \) twice as much as \( Y \)—specifically, if \( p/2 \) probability of \( X \) makes the same contribution to the value of a lottery as \( p \) of \( Y \).

This way of interpreting EUM implies that my basic ‘ought’ judgments logically entail lottery preferences that comply with EUM. So my preference ordering is a representation of a logical fact about what judgments spawned from my desires, taken together, recommend. The same is true of my utility function. My preference ordering and utility function are about my practical stance—the collection of judgments I make on the basis of my strengths of desire and what they recommend.

One reason utilities and preferences are not about desires themselves is that my desires do not obey the completeness axiom but my judgments about actions do. For example, I may not have any desires one way or the other about being a professional curler if I have never thought about it before. But a judgment that I am under some degree of positive pressure to do something else will conceptually entail that I have a preference for it over being a professional curler, as I have thus far determined zero pressure on myself to pursue curling. I will take up this argument in greater detail below when arguing for a realist interpretation of completeness.

### 4.2.2 Outcomes

The formal outcomes in EUM are the objects of desire. This means that the rule for how to differentiate outcomes is in terms of how the agent thinks of the options as differentiated. This is often described in the literature as in terms of the ‘preferences that the agent actually has’, rather than in terms of when it is rational, or justifiable, to have a preference between two options.

This has consequences for how we model Oedipus. Oedipus takes one event, sleeping with Jocasta, and thinks of it as two different possible objects of desire—only one of which he actually desires. This means that we should differentiate this as two outcomes in modeling him, one as sleeping with the queen, one as sleeping with his mother. This also helps us reflect more accurately what actually went wrong.

---

3The notion of ‘same contribution to the value of a lottery’ will be formally defined below in the context of the \( v(\phi) \) function.

4See (Pettit 1991), (Broome 1993), and (Dreier 2005).
for Oedipus in this situation. Oedipus isn’t making a mistake in his means-ends reasoning, or, more specifically, in how he goes about getting what he wants. Rather, he is making a mistake in his attempt to represent the way the world is. This doxastic error has implications on the outcome of his practical reasoning process, even though there are no flaws in that latter stage of the process.

As far as EUM is concerned, Oedipus is perfectly rational. He takes the means to his ends as he understands them. This is different from an approach where we differentiate outcomes on the basis of whether one could justify a preference between them, for example, as these are both the same event.

Decision theories resembling EUM simply are not equipped to describe the norms that go into determining whether two purported states of the world are actually one and the same. But that’s not a problem. As long as we are comfortable separating doxastic and practical rationality as two different but interconnected functional processes, EUM can represent the practical process even though the practical process might be negatively impacted by failures of the doxastic process.

4.2.3 The Axioms and/or the Idea of Maximizing Utility

The idea that we ought to comply with the axioms or maximize expected utility will depend upon the content of the ‘ought’ judgments that desires call upon us to make and the teleological functional role of intentions. The collection of basic ‘ought’ judgments that we form will logically entail a set of all-things-considered desires—the thing (or set of things) that an agent takes himself to be under the most pressure to do at that time. Acting is, then, not a logical consequence of the content of these judgments, but the normative response to these judgments by way of intentions.

It is not necessarily the case that the basic ‘ought’ judgments that we form are about compliance with the axioms. In the next section I endorse a position according to which these basic judgments are not about the axioms, but are a certain kind of implication of these judgments. There are at least two kinds of benefit to a position like this. Firstly, there is one kind of mental process that explains the presence of all...

---

5There may be practical pressure to be doxastically accurate in your representation of the world. For example, if Oedipus doesn’t want to sleep with his mother, then he may be under practical pressure to make sure that the woman he is about to sleep with doesn’t fit that description. I will leave out a detailed analysis of this kind of process. I do not think it substantively impacts the project here.

6This is like the teleological functional delineation of a football team into offense and defense. We can talk about and model the role of the offense—talking about its successes and failures—reducing the activity of the defense to an input into our model of the offense. So, for example, we can talk about the offense starting on the 40 yard line after the defense exits the field.
the axioms. Because my basic ought judgments are about a degree of pressure, all the axioms follow from this alone. The alternative would be that somehow each of the axioms arises as a norm independently of the others, which it would be difficult to explain. Secondly, because the axioms themselves are not the content of our most basic ’ought’ judgments, rationality does not require an individual agent to be an advanced mathematician. While any EUM interpretation will imply that an agent abides by the axioms, demonstrating the relationship between the axioms and the various types of EUM behavior is not a trivial task. It would be a positive thing, from the perspective of theory choice, if there was a basic normative judgment from which compliance with the axioms and compliance with expected utility maximization were both obvious and within the reach of everyday people. I believe that we can get closer to this goal by talking in terms of degrees of pressure to do something—or how good an incremental chance at an outcome would be.

4.3 Basic ‘Ought’ Judgments

The collection of judgments that form an agent’s practical stance must be the kind of thing that he can reason from to conclusions about what he ought to do. As we saw in the second chapter, this requires a set of initial “basic” judgments with ‘ought’ content so there is a possibility of a reasoning process that passes the ‘So What?’ test. But what are these judgments like more specifically, and how do they lead to compliance with a decision theory?

I also argued that in order for a decision theory to present a full rational explanation, preference orderings, and the utilities they imply, needed to be about the implications of formal desires—dispositions to spawn judgments of the form “ω is a good thing for me to bring about”, or “I ought to do perform actions that bring about ω”. I take it to be a conceptual truth that if an agent judges something to be better than something else then he takes it to be more important to him to bring it about. This is a kind of extension of the concept of a formal notion of desire that allows for comparisons. Comparisons like this imply that we can modulate our level of commitment to different outcomes—something that I argued that we do by way of strengths of desire. Given strengths of desire, and the conceptual connection between goodness and how important it is to bring something about, we can think of the qualified form of the above judgments as something like “ω is a good-to-degree-U thing for me to bring about”, or “It is important to me to degree U that I bring about ω”.

Of course, it doesn’t appear that desires of different strengths just create a number in my head corresponding to U here. I can’t introspect on my desires and discover
that I like apples to degree 5. So I want to give more detail in this section on what
the content of these ‘ought’ judgments is actually like. This will also allow us to get
a bit more precise about what a strength of desire is independently of the content
of the ‘ought’ judgments it recommends. A graded desire with a particular strength
will be understood just as a TF disposition to form a judgment with a particular
graded ‘ought’ content.

But there are multiple ways that a commitment to bringing something about
could be less than absolute. That is, there are multiple things an agent might mean
when they say that it is more important to them that they bring about \( \omega_1 \) than
\( \omega_2 \). EUM and a competitor, Risk-Weighted Expected Utility Theory, correspond
to different ways of answering this question. In this chapter I will detail the EUM
interpretation. In chapter 5 I will contrast it with the REU interpretation.

For the EUM agent, the introduction of lotteries does not change what pressure
there is to bring about outcomes, or how the pressure to bring about one outcome
weighs against the pressure to bring about another. Here’s an intuitive way reason-
ing could work that would explain why this happens: The agent reasons that the
probability a lottery assigns to an outcome is just a measure of how effective that
lottery would be at addressing the pressure arising from his desire to bring about
that outcome. The more pressure to bring about an outcome (the stronger the desire
for that outcome), the more valuable a fixed degree of efficacy at obtaining it. This
leads to the following idea: probability \((p, \omega_a)\) is just as good at getting me \(\omega_a\) as
\((p, \omega_b)\) is at getting me \(\omega_b\). So if it is indeed more important to me that I get \(\omega_a\),
and there are no other factors modulating deontic normative pressure, then I would
rather have that \(p\) shot at \(\omega_a\).

In order to get more precise about what this means, we need the following formal
definition:

A partition of a lottery \(\Phi_l\) is a set whose members, \(\phi_{li} \in \Phi_l\) are lists of changes in
probability associated with outcomes such that
\[
\sum_{i=1}^{N} (p_{1,\phi_{li}}, ..., p_{N,\phi_{li}}) = (p_1, ..., p_N).
\]
In other words, if you sum the all the changes in probability associated by members
of the partition with each outcome, you will get the probability \(l\) assigns to that
outcome. Partitions also have the following features and representation:

1. A partition \(\Phi_2\) can also be made of a member of another partition \(\phi_{l1}\) such that
\[
\sum_{i=1}^{N} (p_{1,\phi_{2i}}, ..., p_{N,\phi_{2i}}) = (p_{1,\phi_{l1}}, ..., p_{N,\phi_{l1}}).
\]
2. For ease of notation, let \((r, \phi)\) represent the member of a partition that obtains from distributing \(r\) across the probabilities \(\phi\) assigns to outcomes. So if \(\phi = (p_1, \omega_1; \ldots; p_n, \omega_n)\) then \((r, \phi) = (rp_1, \omega_1; \ldots; rp_n, \omega_n)\).

There are an infinite number of partitions that you can construct from any lottery. For example, take the simple lottery \(l_1 = (p, \omega_1; (1 - p), \omega_2)\). You can construct partitions by dividing a lottery into equal parts: \(\phi_{l11} = (p/2, \omega_1; (1 - p)/2, \omega_2)\) and \(\phi_{l12} = (p/2, \omega_1; (1 - p)/2, \omega_2)\). You can construct partitions by assigning zero change in probability to one outcome in some members of the partition: \(\phi_{l11} = (0, \omega_1; (1 - p), \omega_2)\) and \(\phi_{l12} = (p, \omega_1; 0, \omega_2)\). You can also assign negative changes in probability in members of a partition: \(\phi_{l11} = (0, \omega_1; 0, \omega_2)\), \(\phi_{l12} = ((p - 1), \omega_1; (1 - p), \omega_2)\) and \(\phi_{l13} = (1, \omega_1; 0, \omega_2)\).

The description above assumes that agents can compare the value of members of partitions, \(\phi \in \Phi\), of lotteries. While an REU theorist might have reservations about this, it isn’t unreasonable on its face. In fact, we talk as if we value partition members, rather than entire lotteries, all the time: “Would you rather have a 50% chance of a t-shirt or a baseball cap?” Or “What would you pay to avoid losing your chance at an exit-row seat on the plane?” If valuing the parts of lotteries is possible, then valuing entire lotteries is a matter of determining the way in which the value of the parts contributes to the value of the whole. It turns out that summation is the method that most naturally reflects the commitment inherent in a strength of desire. There are at least two ways to see this.

The first is that fixed degrees of commitment to members of partitions is just a restatement of the Independence Axiom. As I mentioned above, valuing \((p, \omega_a)\) over \((p, \omega_b)\) implies that I would rather have a \(p\) shot at \(\omega_a\) over a \(p\) shot at \(\omega_b\). Not only that, but I judge those opportunities to have a fixed relative efficacy at satisfying my desires. So my judgment about the relative contributing value of a \(p\) shot at \(\omega_a\) versus a \(p\) shot at \(\omega_b\) doesn’t change depending on the lottery in which they are embedded.

The second reason one might think the value of a lottery is the sum of the value of its parts is that other intuitive aggregation functions would require an agent to have access to resources beyond the psychological picture presented in the previous chapter. Two of the points from the previous chapter are relevant here. Firstly, desires issue judgments with ‘ought’ content that capture a degree of commitment. Secondly, beliefs about the likelihood of outcomes and these ‘ought’ judgments logically entail claims about what I ought to do. This means that whatever the content of the ‘ought’ judgment I make as a consequence of desire, it must be about probabilities somehow for the logical entailment to hold. But a strength of desire is just a graded degree of commitment; it is simple enough to be represented by just one
number. This seems to suggest that a strength of desire is a kind of value for efficacy at obtaining an outcome.

But if my desires for outcomes give me a judgment about the value of efficacy at obtaining an outcome, how do those judgments combine to give the the value of an entire lottery? I claim that summation isn’t just the most straightforward option here—it’s the only option that respects the independence of value contribution judgments. If I conclude that there is a certain contributing value for efficacy at obtaining a particular outcome, that judgment is made from a desire that regards only that outcome—it is a judgment that is made in isolation from the other possible outcomes of a lottery. So I wouldn’t, for example, multiply the contributing values of members of partitions by one another to get the overall value of the lottery. This would imply that the judgment I made about one outcome had some kind of relevance to the contributing value of other outcomes, which it does not. As long as I am able to assess the contributing value of members of partitions independently of one another, then summation is my only way to combine those judgments while respecting that independence and without additional information. But on my picture, the agent has already exhausted all the considerations relevant to what he ought to do, so there is no additional relevant information.

All this suggests that on EUM, locutions like ‘how much I want ω’, ‘good-to-degree-U for me to do’, or ‘important to me to degree U that I bring about ω’, analyze to the following:

The EUM judgment recommended by a graded desire for ω: To me (subjectively), for me (as grounds for my own action), what is the value contribution from efficacy at bringing about ω relative to efficacy at bringing about other outcomes?

In order to make this informal description perfectly precise, I will formally define the value contribution of a member of a lottery partition below using the v(φ) function. Essentially, it means that the value of the members of a partition of a lottery can be valued independently and compared to find the relative value of the overall lottery.

Efficacy at bringing about an outcome just is the probability at which your action will bring about that outcome. So if I have two levels of strength of desire, more for ωa than for ωb, my desires recommend judgments assigning levels of value contribution assignment.

7The same should hold for the contributing value of some probability shot at one outcome with another probability shot at that same outcome. So, for example, if there is a lottery with a 50% chance of ω, the first 30% has a contributing value to the lottery, and the second 20% has a contributing value, and they do not influence one another.
to efficacy at each, where a fixed level of efficacy at \( \omega_a \) is has a constant relative value to that same level of efficacy at \( \omega_b \).

This judgment operates as a judgment about the strength of a contributing reason to do something, or, a degree of pressure to do it. What this means in a formal sense, and why it implies EUM, I will discuss below.

4.3.1 The \( v(\phi) \) Function and No Interaction Effects

Certainly, my strengths of desire do not give me a judgment about a number corresponding to an outcome. This means that I have no number to multiply by a probability to get the value contribution from \( \phi \). So if my ‘ought’ judgments are about utilities in some non-numerical sense, then how does this imply EUM? Another way of putting this question is how the informal idea of a value associated with efficacy at an outcome translates into the formal components of EUM.

So we know that if one outcome is better than another then a fixed probability at it is more valuable than that same probability at an alternative. It also seems reasonable to think that a higher probability at getting an outcome should create a contributing reason to act that is closer to the weight of the reason following from the sure shot at that same outcome. I mention these considerations to motivate the following: it seems plausible to think that judgments about value contribution from partition members change continuously with strengths of desire and/or probability. We should be able to think of this value for efficacy as represented by a real number line: \( v(\phi) \rightarrow \mathbb{R} \).

The \( v(\phi) \) function appropriate to EUM will abide by the following axioms:

- **Definition of Preference:** For any two lotteries, \( l_1 \) and \( l_2 \), \( l_1 \succeq l_2 \) if and only if \( v(l_1) \geq v(l_2) \).

- **No Interaction Effects:** for any set of outcomes, \( \omega_A...\omega_N \), and corresponding changes in probability, \( p_1...p_N \), \( v(p_1, \omega_A; ...p_N, \omega_N) = v(p_1, \omega_A) + ... + v(p_N, \omega_N) \).
  
  1. This holds for all \( \omega_A...\omega_N \), including when outcomes are repeated—such as when \( \omega_A = \omega_N \). So, for example, for any \( p \) and \( \omega_A \), \( v(p, \omega_A) = x(p/x, \omega_A) \) by NIE.

- \( v(\phi_{NULL}) = v(0, \phi) = 0 \).

The definition of preference is fairly straightforward. \( v(l_1) = v(\phi_{l1}) \) where \( \phi_{l1} \) is the only member in the partition of \( l_1 \). If it contributes more value than the alternative \( \phi_{l2} \), then \( l_1 \succeq l_2 \).
No Interaction Effects (NIE) does work that is very similar to the Expected Utility Theorem of EUM. With these components, the \( v(\phi) \) function implies all the axioms of expected utility theory. I will explain why, both formally and informally, in the following subsections. The Expected Utility Theorem also implies the components of the \( v(\phi) \) function, but this is more obvious. The difference between them is in what feature of our basic judgments NIE and the Expected Utility Theorem highlight.

The Expected Utility Theorem, \( l_1 \succeq l_2 \text{ iff } \sum_{i=1}^{N} p_i u(\omega_i) \), treats the value of outcomes as a weight we place on levels of efficacy. In other words—the value of a \( p \) chance of something has a linear relationship with the value of the outcome it brings about.

The benefit of the Expected Utility Theorem presentation is that it clearly presents the idea of a fixed relative value for a level of efficacy that is described in our informal description of basic ‘ought’ judgments. But there are two problems. The first is that it doesn’t, on its face, present the reasons why the contributing value of a partition increases linearly with an increase in probability. That is, why does it have to be the case that \( v(2p, \phi) = 2v(p, \phi) \)? This is a mathematical implication of expected utility theory, but the motivation for that requirement is lost on this presentation. The second problem is that it is not obvious how to use it to make decisions. My strengths of desire do not give me an actual number to work with in reasoning about what I ought to do. So I’m not in a position to introspectively calculate expected utility in any obvious way.

NIE represents another way of thinking about the value of a lottery. The benefit of NIE over the Expected Utility Theorem is that it highlights how the structure of lotteries in terms of parts and wholes contributes to the calculation of their value. NIE doesn’t, on its face, say that the contribution from an outcome at a particular probability is the value of the outcome multiplied by the probability, or any other particular function. It just says that the value of the overall lottery is the sum of the value of its parts, where ‘parts’ is understood in terms of possible partitions. Summing is appropriate under this interpretation for the reasons I mentioned above. This implies that a basic ‘ought’ judgment is a judgment about the strength of a contributing reason—that the sum of them is the total weight of reasons in favor of acting.

But just as I don’t have a numerical value to multiply, I don’t have a numerical value to sum either. But NIE implies another way of determining whether one lottery is more valuable than another. All you need to do is figure out whether the parts of one lottery are, on average, or in total, more valuable than the parts of another lottery. In order to do this, you don’t actually need to employ utilities as numbers. If both lotteries are partitioned into exhaustive and easily comparable parts, then
you can see which lottery is made up of the most valuable parts.

In fact, suppose that we construct a partition of \( l_1 \) whereby every member of that partition is deemed more valuable than its partner in the partition of \( l_2 \). So there is an exhaustive pairing of members of the partitions such that for each pair the member from \( l_1 \) is deemed more valuable. It seems intuitive that a relationship between partitions like this would be sufficient for requiring a preference of \( l_1 \) over \( l_2 \). But this is only the case on EUM, and not, for example, on REU.

We can see this in that REU does not require agents to obey Independence, and any Independence violation also violates this condition. Independence is just the idea that if \( l_1 \succ l_2 \), then \((p, l_1; 1 - p, l_3) \succ (p, l_2; 1 - p, l_3)\). \( v(p, l_1) > v(p, l_2) \), so \( v(p, l_1; 1 - p, l_3) > v(p, l_2; 1 - p, l_3) \) by NIE. So \( v(p, l_1; 1 - p, l_3) > v(p, l_2; 1 - p, l_3) \) by the Definition of Preference, keeping in mind that the \( v(\phi) \) function can take entire lotteries as inputs by taking partitions with only one member.

Because of the flexibility of partitioning, if \( l_1 \succ l_2 \), the \( v(\phi) \) function implies that it will always be possible to construct a partition where all the members of the partition of \( l_1 \) are more valuable than corresponding members of a partition of \( l_2 \).

Because \( v(\phi) \) and the Expected Utility Theorem imply one another, the formal representation of fixed relative efficacy at obtaining an outcome can be understood in terms of weighting probability by values of outcomes (the expected utility theorem), or in terms of the value contribution from bits of probability of different outcomes (NIE). Either process amounts to the same thing in the end. The content of our most basic ‘ought’ judgments is represented by the structure of these functions, but is also informally captured by this natural language description.

Actual people may break down lotteries into their component parts and check for favorable tradeoffs. This would be the reasoning process most clearly reflected by NIE. This could be done by finding the difference in probability that two lotteries assign to outcomes and checking whether the leftover partition member has positive value. It could also be done by pairing up partition members of equal value from each lottery and then seeing which lottery is recommended by the final comparison. This second process is very natural. Often we carve off and ignore the components of lotteries that are of equal value only to compare the component in which they differ. For example, if two different lotteries each give me a chance at a different prize relative to the status quo, I might ignore the status quo and just compare the value of the two prizes.

Or, more closely resembling the Expected Utility Theorem, a person might think most directly in terms of utilities as weights associated with probabilities. The point

\[ \text{One way to do this is if each member of the relevant partitions is just the original lottery with the probabilities associated with outcomes multiplied by } \frac{1}{n}. \text{ There is a proof in the appendix.} \]
here is not that they must think in one of these ways, but that whatever process
they use must be functionally equivalent to this process. The real, substantive,
psychological claim I am making is that endorsing a desire with a particular strength
normatively recommends assenting to a judgment about a fixed level of importance
associated with a given probability of the object desired. Or, in other words, that
a strength of desire just is a disposition, in the TF aims sense, to form judgments
about $v(\phi)$ values for a particular outcome. NIE and the Expected Utility Theorem
are just two formal representations of this$^9$.

There is an interesting additional benefit to thinking of basic ‘ought’ judgments
in terms of a fixed level of importance given to efficacy over thinking in terms of
the axioms or any particular formal representation itself. This is that it is much
easier, I think, for an average human being to recognize the appeal of the axioms
concurrently with the appeal of the Expected Utility Theorem if they are thinking in
terms of fixed values for efficacy. If the axioms were somehow the most fundamental
considerations in our minds—as if there was a file in our heads that contained a
list of standards we must not violate—then it certainly wouldn’t be obvious that
maximizing expected utility is a good idea. The conceptual connection between
the axioms and the Expected Utility Theorem is outside the mathematical skill
level of most people. And yet, somehow, the idea gains a great deal of acceptance
independently of people being walked through the proof from the axioms. This
common sense view implies that this happens because people have an intuitive sense
of utilities as degrees of goodness or strength of desire, and an intuitive sense of
how these things exclusively determine how important efficacy at bringing about
outcomes is. On a view like this, the axioms are themselves attractive because they
abide by these rules. It is interesting to note that often in conversation, when people
try to explain the appeal of independence, they talk about $(p, l_3)$ being “just as
valuable” regardless of the lottery in which it is embedded. This suggests that the
axiom is not just a basic concept from our minds written down, but is the implication
of valuing members of partitions.

All I hope for at this point is that this picture tracks what people might take to
be the naive view about how decision theory and reasoning works—a view which I
happen to think is correct. None of this is meant to be an argument against alter-

---

$^9$Another way one might think to represent a fixed level of importance of efficacy at an outcome
is through Fixed Tradeoff Rates. Let $(p, \omega)$ be a member of a partition that assigns a change in
probability $p$ to $\omega$ and zero change in probability to all other outcomes. A fixed tradeoff rate, $r/s$,
exists between $\omega_A$ and $\omega_B$ just in case, $(s, \omega_A)$ makes the same contribution to the value of a gamble
as $(r, \omega_B)$. Fixed tradeoff rates are implied by the $v(\phi)$ function, and the proof is in the appendix.
The trouble with using fixed tradeoff rates as the formal representation of basic ought judgments
is that $r/s$ is undefined in certain cases.
native views, it is just an attempt to present an account. The surface plausibility of this view is enough, I think, to suggest that people have the capacity to reason in this way. The hope is that this surface level plausibility, plus the fact that it does imply EUM, mean that a complete argument in its favor only now requires demonstrating the the axioms follow from these kinds of judgments, and then defending the view against alternatives in following chapters.

4.4 A Rational Justification for Compliance with EUM Axioms

In this section I will demonstrate that the axioms follow from the definition of the \( v(\phi) \) function. This will be enough to show that the theory presented here implies compliance with EUM. But I also want to do some additional work showing why the axioms are intuitive. I’ll do this by arguing that the axioms also follow from a natural understanding of another concept that we employ in reasoning—efficacy. In the following chapters, as we transition to address alternative theories of rationality, it will become increasingly important to test the implications of those theories for the network of concepts surrounding rationality—concepts like efficacy, representative decision making, and justifying our actions to other people. I want to begin that process here by situating the axioms as implications of the \( v(\phi) \) function, and also as intimately connected with another notion of what makes for ideal decisions.

The picture of practical rationality presented in the previous chapter was instrumentalist. This means that practical reasoning is nothing more than an agent coming to be attracted to an object of desire (a way for the world to be) and then acting in compliance with one-directional normative pressure to discover and execute the means to its attainment. This contrasts with theories where practical rationality involves other types of norms. These alternative theories will be the subject of the next chapter. But insofar as we are interested in explicating instrumental rationality, the following principle seems appropriate:

**Constraint on Instrumental Rationality (CIR):** A rational agent cannot choose an option that is, by his own lights, less effective than some available alternative at satisfying his desires.

CIR seems self-evident. The problem is that theorists may disagree about what it means to be *most effective at satisfying* one’s desires. EUM and REU correspond to different ways of answering that question. In this chapter I will only talk about
the EUM notion of efficacy, leaving the REU notion until chapter 5. In this chapter, I will argue that an intuitive version of CIR applies to the EUM axioms.

This is important because the axioms are not undeniably attractive themselves, nor are they in some sense the basic constituents of our mental processes. Rather, our beliefs and desires most fundamentally drive judgments about efficacy that themselves are conceptually equivalent to EUM. As the axioms are themselves also, in combination, equivalent to EUM, we can think of deviation from the axioms as kind of deviation from the notion of efficacy that I argued is employed by EUM agents.

I am not sure why some people find the axioms themselves directly intuitive, and some not, or why people might find EUM intuitive, and some not. One thing to note is that the axioms are connected with certain common types of inferences and applications of our notion of efficacy and instrumental rationality. That is, individual decision problems might most closely resemble a question of transitivity, or a question of independence. Even if our most basic judgments regarding comparative value do not mention those axioms directly, their immediate applicability might make them, for some, more obvious.

Holding fixed that CIR is true, I will go through the Von Neumann-Morgenstern axioms one by one and demonstrate how, firstly, they are implied by the \( v(\phi) \) function or the interpretation behind it, and then secondly, how deviation from any of them leads to an alien and implausible notion of efficacy.

### 4.4.1 Transitivity

Transitivity - for any three lotteries \( l_1, l_2, l_3 \in L \), if \( l_1 \succeq l_2 \) and \( l_2 \succeq l_3 \), then \( l_1 \succeq l_3 \).

Transitivity follows from the definition of preference in the \( v(\phi) \) function. If the sum of the partition of \( l_1 \) is greater than that of \( l_2 \), then the sum of the partition of \( l_1 \) will be greater than that of \( l_3 \).

The strangeness of violating transitivity on an instrumentalist picture can be seen in what it means for ‘efficacy’ in CIR. Normally, we take efficacy itself to be transitive. If \( A \) is a more effective than \( B \), and \( B \) is more effective than \( C \), then \( A \) must be more effective than \( C \), assuming \( A, B, \) and \( C \) all apply in the same context to the same type of problem. Violating transitivity and remaining an instrumentalist would mean that an agent must take it that ‘being more effective at satisfying his desires’ is not transitive. Imagining a case like this, it becomes difficult to understand what the person is actually saying.

In the case of choices between sure shots, this is obviously incompatible with strengths of desire. If I desire \( \omega_1 \) more than \( \omega_2 \) more than \( \omega_3 \), then, of that set, it just must be the case that acting to get \( \omega_1 \) is most effective at satisfying my desires.
The plausibility of transitivity in effectiveness over lotteries comes partly from fixed values for efficacy, but it also comes from the transitivity we see in the concept of efficacy applied in all other contexts. For example, if I judge that car A is more effective racecar than car B, and B more than C, then A must be more effective than C (where 'more effective racecar' means that it gives a preferable distribution of wins). If someone disagrees with that, it is more likely than not that efficacy just isn’t the concept governing their comparison of the cars.

One might object that winning races is a singular goal, so it is much easier to tell when a racecar is more effective than another. Alternatively, lotteries have different probabilities of satisfying completely different desires, and as such there is no single property that they have which makes one clearly more effective than another.

I would say that there is a single property that makes one lottery more effective than another, which is a property about how it makes favorable tradeoffs around efficacy at obtaining outcomes. But leaving that aside, we can separate the implications of a judgment of efficacy from the reasons for it. In some cases, a multitude of different considerations may come into a judgment of efficacy—for example, I may think that such and such is the most effective tax policy, or justice system, etc., even though those judgments are based upon a comparison of many disparate kinds of value. But once I make a judgment of a level of efficacy, I commit myself to placing that policy in a transitive ranking.

Another way to see that efficacy is transitive is that the efficacy of each lottery, choice, car, etc., is independent of the efficacy of the others. The efficacy of a car in a race is determined by facts that are independent of any considerations about an alternative car I might use. So suppose I select car A from a choice between all three options. Then, in a choice between just two, A should still come out on top because the considerations that made it more effective than the alternatives do not supervene on the menu from which it is chosen.

This point about supervenience is essential, and it is also what governs transitivity in case of desire satisfaction. Because outcomes are differentiated to take any menu sensitivity into account, judgments of the efficacy of actions also do not change given changes in the properties of menus themselves. So if I judge A to be preferable to two options at once on the basis of how it satisfies my desires, changing the choice to being just between A and B will not introduce any facts on the basis of which CIR and the content of rational deliberation demand that an agent makes judgments that imply that he is not being less effective than he otherwise could be. So even if an agent can act rationally without an explicit judgment about efficacy, he will still be required to act within the constraints of such a judgment, compatible with his desires, were he to explicitly reflect upon it. In such a case, the judgments would be implicit in the sense mentioned in chapter 2.
the relative efficacy of the actions at satisfying my desires will change. This implies transitivity because it creates a stable ranking of alternatives.

4.4.2 Completeness

Completeness - for any two lotteries \( l_1, l_2 \in L \), \( l_1 \succeq l_2 \), \( l_2 \succeq l_1 \), or both \( l_1 \sim l_2 \).

As long as the \( v(\phi) \) function is always defined, completeness will follow obviously from its definition. But it isn’t obvious why completeness should follow from the interpretation I’ve been presenting—that is, why the \( v(\phi) \) function should always be defined. After all, why think I have desires regarding every possible outcome? If someone asked me, for example, how I’d like being a professional curler, I’d have to admit that I hadn’t thought much about it, and as such hadn’t formed a clear desire one way or the other.

Most outcomes instantiate some combination of properties that I either desire or actively disprefer. For example, a career as a professional curler involves a certain expectation of pay, which is important to me. And at the fundamental level, it would be difficult to find options that did not at least have implications for how much pleasure or pain I might feel as a result—both properties that I likely have a preexisting disposition about. In cases like these, my desires place pressure on me to figure out which option best satisfies my desires as that is simply part of the process of taking the most effective means to satisfying them. So, in that sense, there is pressure to have a preference.

Other outcomes might have no properties that are relevant to any of my desires. In such a case, I am still under pressure to have preferences over them. I should actively disprefer them to options that on balance satisfy my desires, and I should prefer them to options that on balance bring about what I am actively averse to. In more formal terms, this means that these sorts of options, at any probability, make the same contribution to the value of a lottery as a null member of a partition. So even if there is not necessarily pressure to have a desire regarding a given outcome, there is pressure to place it in a complete preference ordering in order to be as effective as possible at satisfying the desires you do actually have.

The strangeness of a violation of completeness on this picture is something between two extremes. On decision theories where preferences are defined in terms of choice behavior, a failure of completeness looks like Buridan’s ass—standing in place starving to death because he has an incomplete preference ordering over two bales of hay. On the other hand, incompleteness in preferences interpreted as psychological states might be taken not to be that bad—it is just the absence of certain comparative attitude. While I do not think incompleteness would cause a person to starve
to death, it is more than just the absence of an attitude. Incompleteness is a kind of irrationality that results from failing to appreciate the implications of the judgments that arise from your desires. If I am under some degree of pressure to do something as a result of a desire, then I am under more pressure to do it than to do something for which I have no desire and that there is no pressure to do. Because there is always some level of pressure or other, even if that level is zero, completeness follows.

Notice that this is a reason to think that utility functions are, most strictly, about an agent’s practical stance, and not about their strengths of desire. The utility function is complete, and so are the implications of the judgments that agents make on the basis of their strengths of desire (their practical stance). But strengths of desire are not complete. There may be outcomes that I simply have no desire about.

One implication of transitivity and completeness in the context of differentiating outcomes to capture menu sensitivity is that it creates what John Broome calls ‘non-practical’ preferences. Nonpractical preferences are preferences between actions that could never appear in a menu together. For example, suppose that Maurice would prefer mountaineering (M) to staying home, (H), staying home to going to Rome, (R), and yet prefers going to Rome over mountaineering: \( M \succ H \succ R \succ M \). When asked why, Maurice explains that staying home over a choice to go mountaineering makes him look like a coward, whereas when he has the alternative of going to Rome, he has a good excuse not to hike up the mountain. So, as the suggestion goes, mountaineering is really two different outcomes corresponding to the different desires that are satisfied given the different menus. We have \( M_H \), mountaineering when the alternative was home, and \( M_R \), mountaineering when the alternative was Rome.

This alleviates the intransitivity, but it causes a different complication. Maurice’s new preference ordering, \( M_H \succ H \succ R \succ M_R \), suggests, given transitivity and completeness, that \( M_H \succ M_R \). But this choice could never happen. This raises the question of what it could mean to have a ‘nonpractical’ preference such as this one. Broome thinks that this is an insurmountable problem for an account of instrumental rationality based on desires because he does not think that agents have access to these kinds of preferences. Buchak thinks nonpractical preferences are unproblematic because she thinks that an agent can take a standpoint that abstracts away from the fact that we can’t actually choose between these options and assess something like how much pressure there is to perform either action.

While I agree with Buchak that you can take this abstracted standpoint, I want to add that only a very specific kind of menu sensitivity leads to preferences that actually have no practical impact. This is because the formal notion of desire that
is necessary for a decision theory to present a full rational explanation is exclusively compatible with sophisticated choice in sequential choice situations, and these ‘non-practical’ preferences will have practical importance in those contexts. Furthermore, these considerations suggest that it is more plausible that staying home, not mountaineering, gets redifferentiated. This is because our story of Maurice attached cowardice, a kind of disvalue, to staying home. This has different practical implications than a differentiation in mountaineering.

Interpreting preferences orderings in terms of a formal notion of desire means that a rational agent will always choose the highest expected utility option when evaluated from a given decision node. This is because all factors, including properties about the structure of the problem itself, have already been included in the agent’s formal desires and expected utility assignment at that node. As mentioned in chapter 1, this is what allows the utility assignment to serve as a complete explanation of why the agent chose what he did. For example, suppose the numbers at the right of Figure 4.1 are my utilities.

At node 1, I know, by hypothesis, that I will prefer $C$ to $D$ and $E$ to $F$. In other words, the content of my formal desires will logically entail that I ought to $C$ or $D$ at that time, depending on the first choice I make. Therefore, I recognize at node 1 that my options are either to set myself along a path to choose $C$ and get 4 utility or choose $E$ and get 3 utility. So, because at node 1 I prefer the outcome down the path of $C$ to that down the path of $E$, I choose $A^{11}$.

The utilities of the outcomes in the terminal decisions have practical implications for the choices I make at earlier branches. This is still true for Maurice, and most cases of purported ‘nonpractical’ preferences. Suppose we differentiate mountaineer-

---

$^{11}$This is sophisticated choice.
Maurice recognizes that he would $C \succ D$, that he would $F \succ E$ and that he would $D \succ F$. So, by transitivity, $C$ would be the most effective way of satisfying his desires, and the most effective way to $C$ is to $A$ at node 1. This is a practical implication of a ‘nonpractical’ preference.

In the case of my description of Maurice—in that he doesn’t really seem like he wants to go mountaineering at all, but only goes because he would feel like a coward if he stayed home instead—it seems more accurate to redifferentiate H rather than M. We can see this in two different ways. The first is in terms of Maurice’s basic desires. He actively wants to not be a coward. This creates an additional failure in desire satisfaction if he stays home rather than going mountaineering that does not occur when choosing between Rome and home. The second way that we can see that this is the best way to redifferentiate the outcomes is through looking at the utilities and the implications for the extensive form decision problem. In the problem above, Maurice ends up going mountaineering. But in the story, it doesn’t really seem like he wants to go mountaineering at all—he just doesn’t want to look like a coward. It isn’t that the situation made mountaineering better than going to Rome, rather, it made staying home as a coward worse than going to Rome. So, really, Maurice’s preferences should look like $H \succ R \succ M \succ H_C$. This leads to the sequential choice problem in Figure 4.3.

Given that he knows that he would choose mountaineering over being a coward and staying at home, he decides that he would rather avoid making that decision and choose between Rome and mountaineering, ending up in Rome.

I do not think that this changes the epistemology of my preferences—I reflect on how good I think it would be (how much I desire) that something happen, and on that
basis I determine what I think I ought to do. This could happen without a practical relationship between ‘nonpractical’ preferences. I bring this up simply to note that most of these preferences are not exactly nonpractical. It is good that there is an epistemology of preference that doesn’t rely upon a practical relationship between all preferences, however, because there are some ways of defining globally sensitive preferences that can’t have any kind of practical relationship. For example, suppose that an agent attached an extremely high premium to $M_H$ when it isn’t part of sequential choice problems. This premium has implications in preference that have no practical impact I can discern. But all other truly nonpractical preferences like this seem equally uncommon.

### 4.4.3 Continuity

Continuity - for any three lotteries $l_1, l_2, l_3 \in L$, if $l_1 \succeq l_2 \succeq l_3$, then there exists a $p \in [0, 1]$ such that the agent is indifferent between a $p$ probability of $l_1$ coupled with a $1 - p$ probability of $l_3$ and getting $l_2$ with probability 1.

Continuity follows from the definition of the $v(\phi)$ function. A proof is in the appendix.

Continuity is different from the other axioms in the sense that it does not have as close a connection with the concept of efficacy. That is, it doesn’t appear that preferences that violate continuity require an alien notion of efficacy. But it just so happens that human desires come with finite strengths relative to one another, and as such preferences are continuous. This is a consequence of the instantiation of the teleologocal functional role of desire in us—not a conceptual requirement of that role, or a conceptual requirement of ‘taking the most effective means’ as such, or even incompatible with the idea that desires are for the purpose of forming ‘ought’
judgments. So the strangeness of violating continuity is revealed in a slightly different way from the other axioms.

Another way of phrasing continuity is that there are no lexicographic preferences, or no infinite utilities. Having lexicographic preferences is not itself incompatible with a common notion of taking effective means. For example, if I had lexicographic preferences, the most effective actions, from my point of view, would be to exhaust every opportunity to increase the probability of satisfying my top ranked desires, no matter how small. Then I would do the same with the next tier of desires, so on and so forth down the line.

But lexicographic preferences themselves are strange enough to suggest that they just don’t happen in real people. Any time a person articulates a preference that appears lexicographic, there are usually implicit limitations that reveal finite utilities. Take the following example. Suppose a general tells a colonel to devote all his resources to defending the coast if at all possible. Then, only after defending the coast has been completed or deemed impossible, he should defend the mountain pass. The literal meaning of these orders seem to reveal lexicographic preferences. And yet, if the colonel, unable to locate the enemy coming by sea, fired all munitions blindly into the ocean in order to obtain what he calculated to be a $10^{-200000}$ chance of successfully defending the coast, the general could reasonably complain that the colonel failed to recognize an implicit standard about the appropriateness of actions as means.

It is a simple matter to come up with other examples that reveal that even though people talk in terms of ‘do one thing at all costs before some other thing’, they don’t really mean that. Actions that are not expected to obtain satisfactory outcomes if repeated every second for the lifetime of the universe, for example, simply happen to be inapplicable as effective means to the kinds of desires human beings happen to have.

One objection to this position comes from the case of moral or religious belief. People influenced by Kant, for example, might think that rational preferences take all moral considerations first, and then, only once the expectation of moral goodness has been maximized, can we act so as to fulfill other aims.

I am not convinced that this is even plausible on its face\textsuperscript{12}. But lexicographic preferences wouldn’t be the right way to capture moral normativity even if this were true. This is because while there should be deontic normative pressure in the immoral agent to be moral, it seems implausible that there would be an internalist source of normative pressure that can get an agent from continuous preferences to lexicographic preferences. So suppose there was an immoral agent with continuous

\textsuperscript{12}(Foot 1972) provides some good reasons why not.
preferences in a world where morality amounts to having lexicographic preferences (and deontic normativity is a matter of the effective expression of one’s preferences). Then there would be no deontic normative pressure for this otherwise reasonable person to be moral. This seems to fail the extensional adequacy success condition for deontic normative considerations.

To flesh this out a bit more, on the view I have been presenting here there is nothing stopping preference change from being rationally motivated. If I am capable of changing my preference by choice or habituation, and the outcome I prefer on my current preferences involves changing the preferences that I currently have, then that is what I will do. Not only is it allowed in theory, but it seems perfectly common. For example, I may discipline or habituate myself to enjoy things that I know are important to my broader goals—goals the success at which requires developing that passion and discipline. Being an artist or writer might be like this. Being happily married might be like this. Moral development might be like this. Etc.

But this model implies a restriction regarding what preference orderings can seem appealing if I start out with continuous preferences. Adopting an infinite utility for something would require a willingness to sacrifice everything else for any chance at that outcome. The expected negative impact of this change on the satisfaction of the rest of my desires will make that preference change unattractive from the perspective of any plausible starting set of non-lexicographic preferences. Adopting morally lexicographic preferences at this point in my life, for example, would require giving up all of my other commitments. And I wouldn’t give up all those commitments just to do good, but to add in any additional chance of doing good. Everything else would be crowded out. I would be the worst version of Susan Wolf’s Moral Saint (Wolf 1982).

The problem is that this cannot seem attractive to me from where I am right now. If I do, in fact, desire the things I desire in the proportion that I desire them, then I will not be willing to take on that probable risk of failing to achieve those things once I have adopted my new preferences. So even if it were possible to adopt those lexicographic preferences, no decision theory in the vicinity of EUM could explain why adopting them was a rational thing to do if I didn’t already have them. In other words, it couldn’t explain the pressure on the immoral agent to be moral.

Even if we could demonstrate that each preference ordering recommends adopting another that is a bit more moral than the one I have currently adopted, that still won’t get me all the way to infinite utilities—it just means that my utility for moral outcomes will continue to grow beyond any finite boundary, but will, at any given point in time, be finite.

It seems to me that what drives the intuition that moral norms might stem
from lexicographic preferences is the idea that moral demands are overriding—that I always ought to do the right thing. But this could truthfully describe a case where selfish non-lexicographic preferences demand adopting more selfless ones—like what might be suggested by a case of aspiring to be happily married, or be a contributing member of society. In these cases what is morally good is plausibly understood to allow room for my own individual interests and pursuits—something that both seems plausible and that lexicographic preferences cannot capture.

4.4.4 Independence

Independence - for any three lotteries \( l_1, l_2, l_3 \in L \), if \( l_1 \succeq l_2 \) and it is not the case that \( l_2 \succ l_1, (l_1 > l_2) \), then for any \( p \in [0, 1) \), \((1 - p, l_1; p, l_3) \succ (1 - p, l_2; p, l_3)\).

Independence follows from the definition of \( v(\phi) \) because the value of the whole can be thought of as the sum of the value of the parts. \((p, l_3)\) makes the same contribution on each side, and \((1 - p, l_1)\) makes a larger contribution than \((1 - p, l_2)\). There is a proof of this in the appendix.

Independence is the most controversial of the EUM axioms, so chapter 5 is dedicated to defending this axiom. But I can make a few preliminary points here about its relation to a natural concept of efficacy and No Interaction Effects.

One thing it might be natural to think, but that deniers of independence would absolutely reject, is that the efficacy of an entire lottery is the sum of the efficacy of the members of its partition. So a \( p \) chance at \( \omega \) has the same degree of efficacy at satisfying your desires, and same contribution to the overall efficacy of a gamble at satisfying your desires, regardless of the other components of the lottery.

A similar supervenience point about efficacy may hold here. An independence denier may reject the idea that we could assess the value or contributing efficacy of a member of a partition taken by itself at all. After all, we could never face a \( p \) change of \( \omega \) in isolation from the rest of a lottery unless \( p = 1 \). So what is my reason for thinking that we can assign a value to members of partitions?

Above I mentioned some reasons that valuing lotteries by their parts was superficially plausible—both because it matched linguistic practice and because it naturally fit the psychological picture presented in the previous chapter. My ultimate support for the idea that we can think of efficacy and value in this way is the result of a confluence of reasons. There is, of course, the idea that it is part of an intuitive and usable folk psychological account of rationality. But primarily, an argument I

---

13 Even if you ultimately do not accept the idea that preferences should be continuous, the worst case scenario here is that the utility function is just an approximation in this respect. (Joyce 1999) argues this position.
will save for chapter 5 is that the implications for denying Independence seem alien enough to suggest that actual people just aren’t like that.

Ultimately defending Independence will require investigating the implications of Independence violations beyond the concept of efficacy itself. In chapter 5 I will look at how Independence seems to be an assumption of the way we represent the decisions of other people, and how we justify ourselves to others. For now I just want to note that Independence follows from the picture I have presented here—a picture where agents break down the components of lotteries and assess their contributing value independently of other parts of the same lottery.

4.5 TF Solutions to the Tension between NR and CR

The objection that representing how people ought to be, (deontic) normative representation, is in tension with representing how they actually are, characterization representation, is touched upon in different ways by a variety of authors\textsuperscript{14}. I want to break down the objection into two versions that capture the spirit of what they are all driving at.

- **Tension between Norms and Characterization 1 (TNC1):** In order for a theory to be an accurate representation of normative pressure on agents, it must be an accurate characterization of their psychologies. But EUM is not an accurate characterization of agent’s psychologies, so it cannot be an accurate representation of the normative pressure upon them.

- **Tension between Norms and Characterization 2 (TNC2):** Suppose that EUM is an accurate characterization of human psychology. If this is true then human beings would always be perfect expected utility maximizers. But human beings aren’t perfect expected utility maximizers, so EUM can’t be an accurate characterization of human psychology.

4.5.1 TNC1 and Reply

*(Meacham and Weisberg 2011)* begin by noting that the justification for thinking that EUM represents human psychology must be either empirical or non-empirical. Their thesis is that if it is empirical it is unjustified and if it is non-empirical it

\textsuperscript{14}See (Korsgaard 2008), (Bermudez 2009), Hulse Read Schroeder, (Meacham and Weisberg 2011)
is uninteresting or vacuous (6). They take behavior that seems incompatible with choosing what I prefer, or most desire, to be an example of a psychological deviation from EUM. This connects well with the approach I have been taking here—trying to paint a picture in terms of desires that will eventually be mapped by EUM.

With respect to the first horn of the dilemma, there are both behavioral and neurological reasons to think that EUM is not empirically justified. Behaviorally, human beings exhibit weakness of will, EUM-inconsistent risk behavior, and non-ideal responses to evidence, just to name a few. If our thesis that human psychology matches EUM is supposed to be empirically justified by observing behavior, then we do not have good reason to think this is true.

Neurologically, the story is just as bad. There are examples that suggest that the structures in the brain are not naturally divided into segments that map onto the objects in EUM/folk psychological description—things like beliefs, desires, etc. Take the following example: A man sets out on a drive to a new job that requires him to turn right at an intersection where he has been turning left for years. Almost automatically, he turns left. It seems odd to say that he had a desire to turn left. However, there are neurological structures that would have formed over all those years of turning left at that intersection that cause turning left at that intersection and that would have persisted even after getting a new job: “In the end, it seems most plausible to take behaviors like that of the driver, behaviors like turning left out of brute operant conditioning mediated by the basal ganglia, as having nothing to do with one’s desires” (Schroeder 2005). Considerations like this convince Hulse, Read, and Schroeder that there really is nothing in the head that corresponds to the notion of conscious desire that EUM seems to present (Hulse, Read, and Schroeder 2004). Neuroscience describes the proximate causes of human action, and it seems to come apart from folk psychological description, so why why think folk psychology gives us any kind of true or meaningful explanation of human mental states as related to the actions we perform?

One way to break down the challenge from TNC1 is into how EUM fails to accurately represent both ‘premise states’ and ‘conclusion states’. Premise states are the input into practical reasoning—purportedly beliefs and desires. But TNC1 says that real human psychological states aren’t like beliefs and desires. So there is no reason to think that the recommendations of beliefs and desires, whatever they might be, apply to real human beings because real human beings don’t have those mental states. Conclusion states are the output of practical reasoning—things like actual choice behavior and statements of preference. EUM doesn’t seem to capture these either because people fail to live up to EUM standards all the time.

I won’t go into the reasons why Meacham and Weisberg think that if EUM is non-
empirical it is uninteresting because I think the thesis that EUM matches human psychology is empirically justified. There are two problems with the argument above. The first is that it doesn’t fully explore two possible sources of empirical justification—introspection and linguistic practice. Secondly, it fails to appreciate the complexity of the relationship between a third-personal and first-personal presentation of the mind.

With respect to leaving out possible sources of empirical justification, notice that there is a very real sense in which you can determine what it is you’re up to when you’re believing something, or trying to satisfy one of your desires. Taking the case of belief, it is very natural to recognize yourself as engaged in a task of trying to represent to yourself the way the world is. That fact may be lost in a description of my behavior only—I may predictably mess up when trying to represent to myself how likely certain states of the world might be, such that giving a description of me as trying to represent likelihoods is difficult using purely behavioral information. But, as it turns out, we don’t need to use only behavioral information because we have introspective access to what we are up to.

Oddly enough, there is a type of behavioral evidence that these objections do not take seriously—linguistic. A piece of evidence for the idea that we have introspective access to beliefs, desires, etc., is the functionality and history of language using these terms. When someone tells me that they believe, or desire, something, I understand what they are trying to tell me. With respect to conclusion states this kind of evidence is less obvious. It comes from the kinds of claims we make about efficacy and making good decisions. The cases I bring up in chapter 5 in defense of the Independence axiom highlight some of the ways of practices reveal a commitment to EUM compatible conclusion states.

Evidence for the existence of colors works in a similar way. If I had to give a theory of color using just neurological and behavioral evidence without linguistic practices, it would be very difficult. Given the multitude of ways in which a visual experience of blue, for example, can arise as a result of the interaction of my eyes with combinations of wavelengths of light, coupled with my inconsistency in picking out colors in different conditions, it might be very natural to conclude from behavioral evidence that there is no such thing as blue. But, if we allow ourselves linguistic

---

15Some of the definitions that my EUM interpretation uses are stipulative. For example, formal desires carve out a teleological functional role that is simply different from an available alternative. This is also true for strengths of desire. While these definitions are stipulative, they do track real properties of practical reasoning. The only reason the definitions are stipulative is because we have to be clear about what feature of the process of deductive reasoning we are trying to talk about. My contention is that this could, to some, conceivably, be interesting.
and introspective evidence, it seems obvious that blue is in fact a property that we have access to. The puzzle is not, “how do we explain all this talk of blue given that it doesn’t exist?”, but “how does this experience of blue relate to neurological and behavioral facts that seem so different from the third-person point of view?”.

This brings us to the second problem with the arguments for TNC1, which is that it fails to appreciate the complexity of the relationship between first-personal mental facts and third-personal neurological facts. Certainly, it might be the case that the brain forms structures that, when viewed from the third-person point of view, would naturally be carved out as identifiable parts of the brain. These parts, when viewed from that perspective, may not map neatly onto the way we would carve up our mental processes from the first-person point of view. But that does not show us that the first-person point of view is an inaccurate characterization of those mental states—rather, it is just a description in different terms. The specific translation of those neurological facts into first-personal mental facts, and why that translation obtains, is the hard problem of consciousness. The empirical data cited in TNC1 is not support for the idea that there are no desires. Rather, it is support for the idea that the answer to the hard problem of consciousness will be more complex than just matching up natural third-personal parts with natural first-personal states.

Also, notice the sense in which characterization of mental states must be ‘accurate’ in order for that characterization to be necessary for capturing normative considerations. As discussed in chapters 1 and 2, the relevant accuracy of characterization is first-personal. The idea is that an agent with these mental states must cognitively apprehend the attractiveness of reasoning in a particular way—this is what drives the universalizability of practical reasoning as a conscious deliberative process. This, as mentioned above, is different from a straightforwardly causal, or third-personal, description of what an agent is up to.

4.5.2 TNC2 and Reply

TNC2 charges that any EUM interpretation mischaracterizes people as always rational. This would be an unacceptable result because it contradicts how people actually behave, and it also undermines the idea that we think of normative theories as po-

16It is also worth noting that parts are different from states, even if you are talking exclusively third-personally. A desire, for example, is a mental state—a state, potentially, of the entire mind/brain system. This is different from what a particular physical part of the brain does. It could be the case that a mental state corresponds to the action of a particular part of the brain. But just because it is difficult to find an individual part of the brain that corresponds to a folk-psychological mental state does not investigate the possibility that the state is instantiated more broadly across the brain.
tentially providing guides or helping us plan actions (Korsgaard). If I’m always going to be rational as a matter of causal necessity, it isn’t clear why using decision theory as a guide would have to be part of that causal story. Presumably, we need guides to guard against a possible deviation from rationality. Compliance by causal necessity rules out that possibility.

But what I hope is relatively clear by this point is that ‘accurate characterization’ doesn’t imply a straightforward causal characterization. The characterization that we need in order for mental states to explain normative pressure is the first-person universalizability of the reasoning process, not universalizability in virtue of causal powers. Teleological Functionalism provides a story for how we can get an accurate characterization that allows for irrationality while still giving a kind of causal explanation of rational action.

4.5.3 Objection: Do we really need to respond to TNC?

One might think that while teleological functionalism and its reliance upon first-personal properties does resolve these objections, these objections are not worth finding solutions to. The premise that they all rely upon is the idea that characterization of mental states is necessary in order to explain normative pressure. But perhaps this is simply not true. Maybe we only need to characterize the mental states of rational agents, or only some of the mental states or irrational agents, in order to have a theory of practical reasoning.

I would claim that TNC objections are actually very much to the point of one of the primary difficulties in explaining deontic normative pressure.

Firstly, they claim that the input into practical reasoning—things like beliefs, desires, etc.—cannot exist in the form that they need to exist in order for EUM to be a plausible representation of what is happening. Secondly, they argue that the conclusion states are nothing like what EUM predicts.

Meeting both of these conditions is necessary in a sense, but it is not the sense that TNC assumes. So we need to do the work of carving these senses apart. The sense in which we need to accurately represent premise states is given to us by the Deontic Appeal and Causal Story success conditions in chapter 1. That is, we need a sense in which an agent is subject to a standard that could be causally efficacious in getting him to do something. So premise states need to be accurately represented to the extent that they can provide a certain kind of first-personal experience and, in the cases of rationality, have a causal impact that resembles the EUM description of that first-personal experience. This is an entirely different kind of standard from the evidence TNC cites, which is about parts of the brain from the third-person point of
view and what they tend to do. In this sense, a theory of rationality does not need to accurately characterize premise states.

With respect to conclusion states, a theory only needs to accurately characterize those in a certain sense as well. It needs to capture why agents who end up acting irrationally can still appreciate the pull of rationality. So the Deontic Appeal success condition still applies. EUM only displays what happens when people are rational because it presents a picture of what people are aiming at (in the TF sense), whether they are rational or irrational. In displaying what happens when things go right, it does not have to present what actual people merely tend to do. Because teleological functionalism characterizes an agent’s practical stance in terms of first-personal properties, it meets the Deontic Appeal condition while allowing that agents systematically fail to meet their own standard—even if they do so for neurobiological reasons.

4.6 Conclusion

In this chapter I presented a view where agents reason in terms of the value of an amount of efficacy at getting an outcome. This leads to compliance with the EUM axioms, and a fairly straightforward realist interpretation of the formal model. What we have is a precise description of what one might have thought EUM was about before studying decision theory extensively. Outcomes represent the objects of desires, and a utility represents a degree to which I think I ought to pursue an outcome. EUM is a good theory because the psychological platitudes that underlie it are well-chosen, and Teleological Functionalism helps us understand how those platitudes could be true in light of superficial dissimilarity with behavioral and neurophysiological data.

These first three chapters present an instrumentalist account of rationality without sufficiently defending it from objections and alternatives views. While there are too many alternatives to address them all, the next two chapters investigate norms of coherence as such and the possibility of violating Independence, respectively. Because the view presented in these first three chapters adheres so closely to the common sense view of what happens in practical reasoning, my hope is to lend it a good deal of support by defending it against these two alternatives.
Chapter 5

Coherentism

5.1 Introduction

The view presented in the past three chapters suggested that deontic normative pressure flows in one direction—that a state, $A$, would normatively require a response, $B$. This was a foundational assumption I used to argue for a picture of rationality where an agent comes to be attracted to an object of desire (a way for the world to be) and then experiences one-directional normative pressure to discover and execute the means to its satisfaction commensurate with the strength at which the object is desired. Positions with this broadly defined structure I will call instrumentalist. Instrumentalist theories of practical rationality have certain benefits. Firstly, they can be expressed in terms of very familiar kinds of mental states and pressures. The idea that we ought to ‘seek the means to our ends’ is one of the least controversial possible norms of rationality. Furthermore, when articulated in terms of the formal notion of desire—as we ought to seek whatever it is that we judge is good to choose—instrumentalism avoids ‘So What?’ objections, as every agent with a judgment with that content will have already appreciated the pressure to bring about the object desired.

A popular strand of alternatives to instrumentalist theories of the normativity of EUM rely upon norms of coherence as such. A norm of coherence as such is a norm such that it would be appropriate to say that it is coherence, not efficacy at achieving an aim, that grounds the universalizability of the norm. These alternative norms can be thought of as existing alongside, or as completely independent of instrumentalist norms. In this latter case, any instrumentalism is merely a redescriptions of what is actually happening at the most fundamental level.

There are two ways that a norm of coherence as such could exist. The first is that
coherence is constitutive of having the kinds of mental states that rational agents have—beliefs, desires, etc.—in such a way as for those norms to be irreducible to, or explanatorily prior to, instrumental norms. This is a way for coherence, rather than efficacy, to become the most fundamental explanation for the universalizability of a norm. In essence, if the set of potentially rational agents are the ones with beliefs, desires, etc., and coherence governs what it is to have those states, the coherence determines why rationality is universalizable. What makes this a norm of coherence as such, is that the coherence, rather than any true description in terms of bringing about ends, is what fundamentally governs the existence of that state. On a view like this, a statement of a true means-ends norm would really just be a description of an effect of coherence norms. For example, if we had reason to think that coherence could explain the appeal of means-ends reasoning, but not vice versa, we would have a reason to think that there are norms of coherence as such.

The second way for there to be a norm of coherence as such is for coherence to be independently and inescapably appealing to all rational agents from the first-person point of view. Another way that authors have described this kind of coherence norm is in terms of what is appealing to the agent about compliance—if I could point out an inconsistency to the agent, he would resolve the inconsistency just because it was inconsistent (Kolodny 2005). In the context of decision theory, this might look like the demand to attend to the axioms (Dreier 1996). The only convenient way to fit this into the psychological picture presented in the previous chapter is to think of coherence as a kind of object of desire (or incoherence as something undesirable), because objects of desire are what the agent finds appealing. In such a case, I may have some sort of overall governing desire for coherence in addition to my other desires, or each desire might come with a caveat about coherence.

Each of the options above is a possible reading of the defenders of norms of coherence as such that I will discuss in this chapter. I will address three different approaches, represented by Davidson, Dreier, and Armendt, for arguing that practical rationality involves norms of coherence as such. But I want to make a preliminary point about why one might want an alternative to instrumentalism in the first place. Remember that strengths of desire are vague. Utility functions demand that I act with unerring precision—there is a difference between a utility of 5 and 5.00001. But if I am tested for transitive preferences around lotteries with small enough probability changes to reveal a difference like that, it is unlikely that I will behave consistently.

---

1One could argue for a position where an entirely new mental state explains the first-person attraction of coherence. However, it is not clear to me that a position like that would resemble a common sense picture of first-person human psychology, and as such it would be implausible on those grounds.
Furthermore, if strengths of desire are vague, it doesn’t even seem true that instrumentalism implies that I ‘ought’ to behave consistently at that level of precision, even though EUM seems to demand it. On the other hand, if desires do not have precise strengths, even if they are only ordinal, additional norms of coherence as such can generate compliance with EUM. One might think that behaving in a way that makes an agent systematically exploitable, or to violate any of the axioms of decision theory itself, is incoherent. If this is true, then even if instrumental norms do not get us all the way to EUM compliance, additional norms of coherence as such could explain the demand to be perfectly precise. So if we are looking for a theory of rationality that grounds the use of EUM as a representation of rational choice, then it seems as though we need to augment instrumental norms with norms of coherence as such.

5.1.1 The Thesis and the Plan

I will argue position that there are no norms of coherence as such and that all rational norms are instrumental. Any true coherence norm is not a norm of coherence as such because it is just a redescription of an instrumentalist norm or norms, which are more fundamental.

I will discuss the three primary methods for arguing that norms of coherence as such are the appropriate normative grounds for EUM. All three arguments fail for slightly different reasons. (Davidson 1963, 1967, and 2004) and (Lewis 1974) discuss a subject matter that is either inappropriate for explaining normative demands, or insufficiently distinct from instrumentalism. (Dreier 1996) seems to make little more than an argument from authority, but which also fails for the reasons Armendt’s argument fails. The third, represented by (Armendt 1993), makes norms of coherence as such seem plausible at first glance. The reason we should not accept Armendt’s conclusion is because these norms end up being a misdiagnosis of a disjunction of instrumental norms.

In the next section I will discuss methodologies for discovering the presence of norms of coherence as such. These methodologies will govern the following three sections where I will go through each of the three arguments for norms of coherence as such and my objections. I will then discuss some takeaways regarding problems with coherence norms generally. In the final section I will discuss an instrumentalist take on exploitability and the impact of vagueness in strengths of desire.
5.2 Detecting the Presence of Norms of Coherence as Such

Keeping in mind that a norm of coherence as such is one where the pressure to be coherent is irreducible to pressure to abide by instrumental norms, there are three sorts of considerations that arise in the arguments I will discuss that will help guide us in detecting the presence of such a norm.

The first way to detect a norm of coherence as such is if coherence is inescapably appealing from the first person point of view. If, in a case of noncompliance with EUM, for example, we find ourselves unable to deny the appeal of the axioms as such, and not merely as redescriptions of maximization or instrumental norms, then we have reason to think we have found a norm of coherence as such. This is most obviously suited to finding norms of coherence as such of the second type—coherence being the fundamental reason why things are done from the first person point of view, not necessarily as a part of the essential constitution of mental states.

The second way to detect whether a norm of coherence as such is present is whether the norm is wide-scope, or bidirectional. When I said that $B$ was the normatively appropriate response to $A$, I implicitly defined normativity in a way that made it one directional—there is really no choice for the agent upon the occurrence of $A$, he has to $B$ on pain of irrationality. One way to represent this is formally, with an ‘ought’ operator. Our narrow scope, or one-directional norm from $A$ to $B$ would be represented as $A \rightarrow O(B)$. This can be read, “If $A$, then you ought to $B$”. But another possibility is that the norm from $A$ to $B$ is wide scope, $O(A \rightarrow B)$, “You ought to (if $A$ then $B$)”. The wide scope norm can be satisfied in two ways. Either $B$, or make it the case that no longer $A^2$. This would be true if any of the normative connections from practical reasoning in our previous chapter were bidirectional. A case where they are all bidirectional is seen in Figure 5.1.

The complication with the terms ‘narrow-scope’ and ‘wide-scope’ is that the same mental process can appear narrow or wide scope depending on the way you carve up mental state functions. For example, take objects of desire, or outcomes. Now, let’s examine norms that connect desires and action:

- **IR:** Always take the most effective means to achieving desired outcomes.

---

2This is not to be confused with an instrumental norm recommending an exclusive disjunction of options. For example, it might be the case that you ought to either $A$ or $B$ exclusively. That can happen because you have a narrow scope desire to $A$ or $B$, and not because there is actually a wide scope norm operating. For example, I might want to be a doctor or a dentist, and this could be expressed as a purely instrumental norm.
• CR: Either take the most effective means to achieving desired outcomes, or stop desiring that outcome.

One norm is narrow-scope, the other is wide-scope, and depending upon what definition of ‘outcome’ you take, either of these could seem plausible. For example, if ‘outcome’ does not include the history that was necessary to reach it, then CR seems much more plausible. For example, if I want to be President, and the only way I can do that is through assassination, then the rational response might be to move on. Alternatively, if ‘outcome’ is taken to include the history that it takes to reach it, then IR seems entirely plausible. The question, then, is whether there is a sense in which the agent really faces a wide-scope norm or narrow-scope norm, or whether wide and narrow scope are just artifacts of ways of describing rationality.

Notice that, using the definition where CR is plausible, CR might actually be a disjunction of narrow scope norms. If CR is truly a coherence norm, then the agent should be free to choose either of the appropriate normative responses presented by the norm independently of any other proximate source of normative pressure. But

---

3This seems to assume I have some control over my ability to mute my desire, or stop desiring altogether. I happen to think that this reflects everyday experience. But if that seems like asking too much, a wide-scope norm here might be described in terms of whether a means is taken or a desire is endorsed—that is, whether an agent is willing to judge that he is under pressure to satisfy this desire.
whether we label it as pressure from a ‘desire’, or as a component of an ‘object of desire’, there seems to be the possibility that, from the point of view of the agent, there is normative pressure not to take certain kinds of means, like assassination. The decision to abandon one’s desire or not is plausibly determined by the competition of two opposed sources of pressure to act—one to achieve what the agent wants, and the other to avoid certain kinds of means. This would imply that the agent is not free to choose in the way that a wide-scope norm would suggest, but, rather, that he is under net pressure to do one or the other exclusively.

So in detecting the presence of a true wide-scope norm, we need to check for this kind of freedom of choice—that the decision between alternatives is not determined by any further normative considerations.

This brings us to our third and final method of testing for norms of coherence as such—untranslatability. Translatability obtains when there is a natural coherence-based description of the deontic requirements in question alongside a natural instrumental description of those same requirements. There are two forms of untranslatability. The first is if the norms are non-overlapping. We have reason to believe in the presence of a norm of coherence as such if the purported norm recommends a behavior, cognitive response, or reasoning process that is different from the behavior demanded by any plausible set of narrow scope norms. The second kind of untranslatability is explanatory priority. Even if a group of coherence norms and a group of instrumental norms recommend the same things, if there is a strong reason to think that the coherence norms explain the instrumental norms and not vice versa, then we have reason to think that there are norms of coherence as such. This is like a case of claiming that chemistry explains biology and not vice versa, even though they are completely overlapping within a given domain. Both versions of untranslatability help discover norms that are irreducible to the demand to be effective at achieving our aims.

---

4Kolodny (2005) makes the point that coherence norms have the problem that you cannot reason from the fact that you must abide by a coherence norm to compliance. I am not sure that this is true. Reasoning is about there being a universalizable cognitive mental process for abiding by the norm. It seems plausible to me that we can, for example, mute, ignore, or abandon our desires if we choose not to pursue them. Additionally, a coherence norm could be universalizable in either of the two ways mentioned above—by being constitutive of mental states or by being independently attractive. Reasoning in terms of a coherence norm means just that my reasoning process allows for the kind of freedom in normative response described here.
5.3 Davidson and Lewis: Radical Interpretation

Imagine that we have undertaken the task of coming to know Karl as a person. We would like to know what he believes, what he desires, what he means, and anything else about him that can be explained in terms of these things (Lewis “Radical Interpretation” p. 331).

Not only are we trying to learn these things about Karl, but we have to do it from scratch—without any preexisting knowledge of his mental states or the language that Karl uses to express any of these things. This is the problem of radical interpretation, a problem that Lewis sums up as “Given P, the facts about Karl as a physical system, solves for the rest” (331).

Lewis is interested in this question not so much because he is interested in doing some anthropology, but because he thinks this problem is a good heuristic for thinking about the problem of how physical facts determine mental facts: “… my problem of radical interpretation is not any real-life task of finding out about Karl’s beliefs, desires, and meanings. I am not really asking how we could determine these facts. Rather: how do the facts determine these facts? By what constraints, and to what extent, does the totality of physical facts about Karl determine what he believes, desires, and means?” (333-334).

A similar view is adopted by Donald Davidson: “What a fully informed interpreter could learn about what a speaker means is all there is to learn; the same goes for what a speaker believes” (315).

Lewis presents a list of six constraints on radical interpretation. He believes that these constraining principles are analytic—that, definitionally, what it is to be a belief, desire, etc., is to conform to the relevant constraints on radical interpretation. However, he admits that he does not need analyticity to justify employing the constraints. As long as the constraining principles “are very firmly built into our system of belief” (336). Or, more directly, that generally abiding by the constraints is somehow a necessary condition of having beliefs, desires, etc., regardless of whether that is a consequence of the definitions of mental state terms or some other fact.

The six constraining principles of radical interpretation are as follows:

1. **The Principle of Charity:** Karl should be represented as having (and actually have) the mental states that he ought to have given the situation he is in and the information available to him.

2. **The Rationalization Principle:** Karl should be represented as (and should actually be) an EUM rational agent.
3. **The Principle of Truthfulness**: Karl usually says and believes the truth, or what we would reasonable expect him to believe is true given his situation.

4. **The Principle of Generativity**: The truth conditions of the sentences in Karl’s language are a function of the constituent parts of those sentences in a relatively uniform way.

5. **The Manifestation Principle**: Karl utters propositions that he actually believes.

6. **The Triangle Principle**: Karl’s beliefs and desires should be the same whether expressed in his language or ours (336-339).

But why? Why think that these constraints are so deeply connected to what it is to have a belief, desire, etc.? It does not seem, on its face, like a conceptual contradiction to suppose that most of Karl’s beliefs are false, that he is irrational most of the time, or that he is predominantly deceptive.

There are two ways of reading what Davidson and Lewis seem to be thinking when they talk about the relevance of constraints on radical interpretation to the content and nature of mental states.

The first reading for why these constraints might matter has to do with the linguistic community and what counts as a possible usable language for describing Karl. The term ‘belief’ itself is supposed to be a term that all speakers of English can, in principle, learn how to use. If you and I are going to communicate about what Karl believes, then we need a shared set of standards about what it means for Karl to believe something. Allowing beliefs to be subject to standards completely inaccessible to us is supposed to make talking about beliefs impossible. This is part of the motivation for the idea that the totality of physical facts about Karl should sufficiently determine what it is that Karl believes, and that we could conceivably figure this out without cutting into Karl’s brain. So what it is to have beliefs must be accessible to an interpreter because it must be possible in the language of the interpreter to competently describe an agent as having those beliefs.

This seems to be the reading of radical interpretation suggested by quotes like “For the only, and therefore unimpeachable, method available to the interpreter automatically puts the speaker’s beliefs in accord with the standards of logic of the interpreter, and hence credits the speaker with plain truths of logic” (Davidson “A Coherence Theory of Truth and Knowledge, p. 316).

Alternatively, one might read the requirement of interpretability as driven by considerations about the nature of content. Perhaps beliefs must be about the events that tend to cause them, and desires must be about the objects that they tend to
cause agents to pursue assuming an EUM framework. This not because it is the only way an interpreter could talk about Karl’s mental states, but because it is a necessary condition of those mental states being about the things that they are supposed to be about. This reading involves understanding beliefs and desires as having intentional properties that are constituted, in part, by the causal relationships they express. This is not a question of definitions and linguistic communities, but a question of the constituent parts of mental states as things, and when those things start to fall apart.

An example that can be read this way comes from Davidson’s statement that everyone, whether they realize it or not, subscribes to the principles of EUM.

This sounds sweeping, even authoritarian, but it comes to no more than this, that it is a condition of having thoughts, judgments, and intentions that the basic standards of rationality have application. The reason is this. Beliefs, intentions, and desires are identified, first, by their causal relations to events and objects in the world, and, second, by their relations to one another. A belief that it is about to rain would lose much of its claim to be just that belief if it did not have some tendency to cause someone who had it and wanted to stay dry to take appropriate action, such as carrying an umbrella. Nor would a belief that it is about to rain plausibly be identified as such if someone who was thought to have that belief also believed that if it rains it pours and did not believe it was about to pour. And so on: these obvious logical relations amongst beliefs; amongst beliefs, desires, and intentions; between beliefs and the world, make beliefs the beliefs they are; therefore they cannot in general lose these relations and remain the same beliefs. Such relations are constitutive of the propositional attitudes (Davidson “Incoherence and Irrationality”, p. 195-196).

This seems, at points, to imply the first reading—if ‘identified’ is taken literally then that seems to be a constraint on the interpreter’s linguistic community. But claims like “such relations are constitutive of the propositional attitudes” seems to make a claim about something that would be true regardless of whether there was a linguistic community outside the agent at all, as a belief would have the same constituent parts regardless of whether there were any interpreters around to talk about those parts.
5.3.1 Radical Interpretation and Norms of Coherence as Such

Both of the readings of Lewis and Davidson imply a kind of pressure towards rationality. On the reading where the linguistic community plays a role, agents are mostly rational because it becomes an impossible task to interpret someone as being overwhelmingly irrational. This is because it is not clear that the terms ‘belief’, ‘desire’, etc. still apply to this person, or are being correctly applied, if the interpreter is attempting to describe the agent as hardly ever doing what maximizes desire satisfaction, or hardly ever believing what is true. An agent can be, at most, occasionally irrational simply as a consequence of the impossibility of using mental state terms in such a way as to interpret him otherwise. On the second reading, we are generally rational because the mental states that we have cease to be what they are when they fall apart. Because beliefs and desires are constituted, in part, by causal relationships, if we do not successfully act and think rationally our beliefs and desires lose their structural integrity—degrading their claim to being beliefs and desires at all. But being practically irrational, for example, is to fail to seek the means to the desires we actually have. If we no longer have desires, then we can’t fail that requirement. Davidson discusses why people must be mostly rational in the following passage:

...it does not make sense to ask, concerning a creature with propositional attitudes, whether that creature is in general rational, whether its attitudes and intentional actions are in accord with the basic standards of rationality. Rationality, in this primitive sense, is a condition of having thoughts at all. The question whether a creature subscribes to the principle of continence, or to the logic of the sentential calculus, or to the principle of total evidence for inductive reasoning, is not an empirical question. For it is only by interpreting a creature as largely in accord with these principles that we can intelligibly attribute propositional attitudes to it, or that we can raise the question whether it is in some respect irrational. We see then that my word subscribe is misleading. Agents can’t decide whether or not to accept the fundamental attributes of rationality: if they are in a position to decide anything, they have those attributes. (It is no doubt for this reason that Aristotle held that an agent could not be habitually akratic; akrasia is deviation from a norm shared by all creatures capable of akratic act) (Davidson, “Incoherence and Irrationality”, p. 196-197).

In addition to applying a kind of pressure towards rationality, both readings of Lewis and Davidson create this pressure by way of norms of coherence as such. In
the case of linguistic practices, coherence is the necessary condition for being able
to ascribe mental state language to a person. In the case of the constituent parts of
mental states, coherence governs what it is to be those mental states. One way to see
this is that, in either case, the agent is free to meet the standard discussed by way of
a wide scope norm that provides for more freedom than a disjunction of narrow scope
norms would. For example, according to Davidson and Lewis, being interpretable
requires that I have mostly transitive preferences, but nothing is included about
revising my preferences in accordance with my strengths of desire. If these coherence
norms are the only norms that explain rationality, then the fact that a rational agent
seeks the means to his ends is secondary to, or explained by, the fact that coherence
is necessary for attributions of rationality.

One might think that part of the constraints on interpretation that must factor
into an account such as this should include a claim about the way in which beliefs
and desires are revised. Such a claim might be, for example, that it is incoherent
to reason from the fact that I am not performing an action to the decision not to
intend that action. In a case like this it would seem that we have a narrow-scope,
rather than a wide-scope, norm operating here. But remember that whether a norm
is wide-scope or narrow scope is only a clue as to whether we are looking at a norm
of coherence as such. The essential property that we are trying to detect is whether
coherence is the root explanation for the universalizability of the norm. It is possible
that we could add a coherence constraint about order of revision and reasoning to
Lewis’s account, which, as long as it was justified by either of our two readings of
Lewis/Davidson, would still be a norm of coherence as such despite admitting of a
true narrow scope description.

5.3.2 My Objections

I have two objections to using the kinds of considerations Lewis and Davidson raise
as justifications for an explanation of normativity grounded in coherence as such.
One objection applies to the first reading of their position, and one objection applies
to the second.

The first objection is that the kind of pressure towards the truth of rationality
claims, or the difficulty in making true irrationality claims, that is articulated by
the first reading of their considerations does not generate the kind of pressure on
conscious deliberation that we are looking for.

The second objection is that on any understanding of the implications of the
second reading for the nature of normativity that captures how we reason, coherence
ceases to be a meaningful explanation for why the norms are the way they are. The
best version of this view ultimately collapses into something similar to teleological functionalism, and as such is perfectly compatible with instrumentalism.

**Objection 1: The Linguistic Community View**

Simply put, the problem with using considerations about the attribution of mental state language as grounds for, or an explanation of, deontic normative pressure is that these considerations are problems for the interpreter, not for the agent himself. We have, for example, no story about how these considerations would even influence the agent’s deliberative process. Facts about what the interpreter is capable of do not influence what seems appealing from the agent’s own point of view (even if they might influence how the interpreter talks about “What is appealing to the agent”), nor do they have any relationship with the operation of the mental states themselves (even if they might influence how we talk about that operation).

If an agent feels inclined to deviate from practical rationality, as in, say, a case of weakness of will, then the fact that he is brushing up against the boundary conditions of interpretability is irrelevant to what he is inclined to do (even if it influences how we talk about what he is “inclined to do”). As such, a piece of advice that we give him about how to act grounded in these kinds of considerations would be completely worthless—it would be a non sequitur from the perspective of his reasoning process.

The way in which coherence needs to be a condition on having beliefs and desires in order for it to provide normative pressure is not as matter of whether a word applies to it, but as a kind of governing rule built into the mind or brain itself.

**Objection 2: The Constitutive Parts View**

The constitutive parts view is at least about the construction of mental states themselves. Coherence does not need to be the agent’s object of desire, rather, in this case it would generate pressure on conscious deliberation obliquely, by making it the case that deliberation must follow the pattern it does because that pattern is coherent, and without coherence mental states fall apart.

The problem is that if coherence is not the object of desire on this view, then why think that coherence is necessary to the construction of these mental states? Notice the similarities between this approach and my own. The idea is that constitutive of having the relevant mental states is a particular tendency towards

---

5 This is a context where the distinction between deontic and non-deontic normativity could be relevant. A linguistic community could employ a standard governed by considerations of language use. But this standard would be a non-deontic norm. They could judge people in light of it, but it would not meet the Deontic Appeal success condition from chapter 1.
rationality—a tendency that can be described using the execution of certain functional roles. The difference is that this tendency is described only in causal/probabilistic coherentist terms for Lewis and Davidson, and teleological functionalist instrumentalist terms on my own view.

I mentioned some of the problems with a purely causal definition of the relationships involved in rational thought in the previous chapter. The idea was that this sort of view will not distinguish between what ought to happen and what merely tends to happen. So we need to investigate ways the Davidson/Lewis position could generate standards independently of how people merely tend to behave. In any case, reasoning is about acquiescence to a kind of first-personal deontic appeal. The question we face here is how coherence as such could explain this appeal. Davidson and Lewis make no claim to coherence itself being the object of desire. So we can investigate two possibilities. In one, the coherence requirement on mental states makes instrumental norms attractive from the first person point of view. Incoherence will make him recognize that he is being less effective at satisfying his aims than he otherwise could be, and that this is a bad thing. In the second, coherence in mental states makes individual norms of coherence as such attractive—for example, the agent will find his nonconformity with the axioms of decision theory unappealing.

In the first case, if the coherence of the underlying mental states makes instrumental norms attractive from the first person point of view, then the rational responses will all be translatable and abide by the stricter narrow scope norms. But this strictness will be illusory. Even though it will appear, from the first person point of view, that there is only one available normative response given my desires, the underlying coherence considerations actually place me under wide scope pressure. So I could, unbeknownst to myself, revise my desires to comply with my actions and still be entirely as I ought to be.

A case like this is implausible just because reasoning comes apart from normative pressure. Normally, we take it that reasoning provides a guide to the normative. In fact, it may be that the structure of first-personal reasoning itself is in some way responsible for normativity being the way it is. So I am willing to reject this version of the Davidson/Lewis position now on these grounds. However, if this is not sufficiently convincing, the objections that I bring up against Armendt and constitutive views of coherence will apply to this position as well.

In the case where individual coherence norms are taken to be attractive to the agent because it is necessary for having beliefs and desires, suppose that the agent was nearing the point of irrationality where his mental states were about to lose their structural integrity. In such a case, the agent is at risk of no longer having beliefs and desires, as defined in causal terms. But that fact—that he is about to no longer
have beliefs and desires understood in that way, is not something that in itself puts any pressure on him. What is so good, after all, about having beliefs and desires like that? It seems entirely possible that an agent could raise a ‘So What?’ objection against the injunction to maintain those states, or his so-called rational status, and the story Lewis and Davidson present does not give any explanation why he would.

It seems, then, that the story that Davidson and Lewis present is not helpful in understanding the nature of normative pressure. Whatever pressure they may have identified related to norms of coherence as such, it does not explain the pressure within irrational agents to make rational choices. This is just another way of saying that they do not provide an answer to the ‘So What?’ question. Teleological functional roles are an appealing alternative to causal descriptions of beliefs and desires because they can explain the sense in which an agent who is not complying is under pressure to comply.

5.4 Dreier: The Founding Spirit of Decision Theory

The description of rationality from the previous chapter generates narrow scope norms dependent upon the content of beliefs about likelihood. This approach is immediately incompatible with (Savage 1954) and (Ramsey 1926) which both offer behavioristic definitions of degree of belief with the consequence that behavior is represented as under wide-scope evaluative (not deontic) norms. (Dreier 1996) and (Dreier 2005) are informed by these approaches and present an interpretation of decision theory as presenting wide-scope coherence constraints on sets of preferences and beliefs. As we will see, Dreier argues that the founding spirit of decision theory is not one where narrow scope norms were the intended subject. One could immediately object that this is an argument from authority, and that whatever the founders of decision theory intended, we are free to use the model in more productive ways now. But it is worth investigating why Ramsey and Savage approached decision theory the way they did in order to check whether our deviation from that method addresses their initial concerns. As it turns out, Ramsey and Savage were operating with an antiquated understanding of mental state terms, an understanding that would leave their position completely incapable of describing the nature of deontic normative

---

6The instrumentalist approach happens to be more compatible with the presentations of (Von Neumann and Morgenstern 1944) and (Mas-Colell, Whinston, and Green 1995) because of their reliance upon the idea that consequences are what matter in rational decision making and degrees of belief are about effectiveness at bringing about various ends.
pressure. Furthermore, even though utility functions are constructed in a way that appears wide scope, they need not be interpreted as implying wide scope norms governing rationality.

In “Truth and Probability” Frank Ramsey attempts to create a system of measuring beliefs, and in doing so, give more precise meaning to the idea that they come in numerical degrees. He is, however, committed to the idea that there is some substantive psychological reality to graded belief, and it is through his system of measurement that we are trying to reveal it (166-167). Savage’s *Foundations of Statistics* is less committal on the matter of the psychological reality of graded belief, and retreats to what he takes to be the safer position “...where a particular interpretation of probability to a person is given in terms of the theory of consistent decision in the face of uncertainty...” (27). The models that they create have a similar structure, and (as far as our aims are concerned) a similar problem. Each defines a particular degree of belief in terms of a particular pattern in betting behavior. As we’ve seen, normative justification is dependent upon the first-personal properties of psychological states. If a “degree of belief” in the model refers to a pattern of behavior, then a “degree of belief” is not what justifies a particular betting behavior. While the results of the model might mimic the behavior of a rational agent, the components of the model are not about the justification of that behavior. This will become clearer by looking at the Ramsey/Savage approach in more detail.

Ramsey’s system of measuring degrees of belief begins with defining *degree of belief* $\frac{1}{2}$.

The subject is said to have belief of degree $\frac{1}{2}$ in such a proposition $p$ if he has no preference between the options (1) $\alpha$ if $p$ is true, $\beta$ if $p$ is false, and (2) $\beta$ if $p$ is true, $\alpha$ if $p$ is false, but has a preference between $\alpha$ and $\beta$ simply. We suppose by an axiom that if this is true of any one pair $\alpha, \beta$, it is true of all such pairs (177).

Because this is supposed to ground a system of measurement, “preference” here is assessed through choice behavior. It is also assumed that $p$ is *ethically neutral*—that it does not affect how the agent prefers the outcomes. He then uses this notion of degree of belief $\frac{1}{2}$ to define sameness of difference in value:

We have to explain what is meant by the difference in value between $\alpha$ and $\beta$ being equal to that between $\gamma$ and $\delta$; and we define this to mean that, if $p$ is an ethically neutral proposition believed to degree $\frac{1}{2}$, the subject has no preference between the options (1) $\alpha$ if $p$ is true, $\delta$ if $p$ is false, and (2) $\beta$ if $p$ is true, $\gamma$ if $p$ is false (178).
These definitions are used in a set of axioms that are then used to correlate real numbers with the values of the alternatives, and then used to define degree of belief:

If the option of $\alpha$ for certain is indifferent with that of $\beta$ if $p$ is true and $\gamma$ if $p$ is false, we can define a subject’s degree of belief in $p$ as the ratio of the difference between $\alpha$ and $\gamma$ to that between $\beta$ and $\gamma$, which we must suppose is the same for all $\alpha$’s, $\beta$’s, and $\gamma$’s that satisfy the conditions. This amounts roughly to defining the degree of belief in $p$ by the odds at which the subject would bet on $p$, the bet being conducted in terms of differences of value as defined (179-180).

I’ve left out the details of the axioms and the proof that lead to the association of numbers representing the values of alternatives. This is because most of the specific details of this proof are unnecessary to the point I am trying to make. Those axioms employ the above definitions, and as such are about the relations in those definitions. These are relations of betting behavior, and as such the conclusion that “This [process of definition] amounts roughly to defining degree of belief in $p$ by the odds at which the subject would bet on $p$” follows because betting behavior is what the axioms are about.

Leonard Savage takes a similar approach in The Foundations of Statistics. For Savage, the relation not more probable than is defined in the following manner:

- Let $\Omega$ be the set of all possible outcomes.
- Let $A, B \subseteq \Omega$ be events.
- $\alpha \succeq \beta$
- The lottery $l_A$ is $\alpha$ given $A$, $\beta$ given $\neg A$
- The lottery $l_B$ is $\alpha$ given $B$, $\beta$ given $\neg B$
- $A$ is not more probable than $B$ if and only if $l_B \succeq l_A$

In the text this sentence captured by the final bullet is ambiguous as to whether the ‘is’ is an ‘is’ of predication or identity. This raises the question as to whether the essential feature of being ‘not more probable than’ is this betting behavior or whether a certain pattern of betting behavior might be something that is just implied by $A$ being not more probable than $B$. At the beginning of the relevant chapter Savage claims to be offering of a definition of probability (27). So it seems safe to assume that the final bullet contains an ‘is’ of identity. If the ‘is’ here is an ‘is’ of identity then
we have a theory that works like Ramsey’s, where probability—or for our purposes a ‘degree of belief in some event’—is defined in terms of betting behavior.

A consequence of the Ramsey/Savage approach is that it does not provide the kinds of explanation of action that we are looking for, or create a model on the basis of which one can give the kind of advice we want to be able to give. The first way to see that this is the case is to recognize that for the presentation of rational explanations, Ramsey/Savage type models are not about the right thing. Remember that a rational explanation is an explanation that justifies a behavior from the first person point of view. In doing so, it answers the ‘So What?’ question and implies that a rational agent will be thus motivated. Suppose then, that all of an agent’s behavior implies a particular set of probabilistically coherent degrees of belief on either of the above models. Then, that agent behaves in a way that is inconsistent with the representation in those models. We have one option: we change our measurement of the agent’s degrees of belief. Strictly speaking this seems like the only correct option to me given the way that degrees of belief are defined on the basis of the choices the agent makes. But suppose, for some reason, we aren’t okay with that approach and we want to assert that the agent actually has incoherent betting behavior. If we go to that agent and say “Look, your degrees of belief are such-and-such, therefore it is rational for you to behave differently than you just did” the agent can respond “So what? I behaved that way in the past but that means nothing to me now”. This is just to play out the worry in (Sen 1973) that started this entire project. In his words, behavior cannot be inconsistent with other behavior, but only with the reason for it. Ramsey/Savage models do not present rational explanations because they are not about the internal psychological states of the agent from which normative pressure issues.

Suppose, then, that one wanted to attach a psychological interpretation to the formal relationships defined in the Ramsey/Savage method. This would amount to treating the ‘is’ in the final bullet of the Savage approach as an ‘is’ of predication.

7The goals of Ramsey and Savage are similar to the goals of Luce and Raiffa. In neither case does the presentation treat us to an explanation of expected utility theory as a set of deontic norms:

“[Certain] (simple) preferences come first and certain rules of consistency are accepted in order to reach decisions between more complicated choices. Given these, it turns out that it is possible to summarize both the preferences and the rules of consistency by means of utilities .... The point is that there is no need to assume, or to philosophize about, the existence of an underlying subjective utility function, for we are not attempting to account for the preferences or the rules of consistency. We only wish to devise a convenient way to represent them” (Luce and Raiffa 1957, p. 32). (As quoted by Dreier 1996).
This seems to be the approach of (Dreier 1996) and (Dreier 2005). There are still two problems with this approach, and they center around the ability of such a model to present rational explanations. Firstly, because of the way Ramsey and Savage both define degrees of belief/probabilities, they do not exist when the agent is indifferent between all options. This is because the necessary lotteries cannot be constructed in order to define degree of belief in either case. Savage is okay with this because he believes such a case is “...too trivial to merit study” (31). It certainly is a boring, and unlikely, scenario, but the content of a psychological degree of belief is conceivably something we might want to attribute to a being that is indifferent between all things. Certainly, by observing the betting behavior of such an individual we would not be able to discern anything about his degrees of belief, but there might be a natural definition of ‘degree of belief’ that captures what is similar about him and an individual that believes what he believes, but actually cares about what happens.

Secondly, notice that the behavioral norms that arises from a psychological interpretation of a Ramsey/Savage model are wide scope. Suppose I have made enough decisions to determine degrees of belief for the following lotteries and I face two choices:

1. $\alpha$ or $\beta$

2. $(p, \alpha; (1 - p), \beta)$ or $(r, \alpha; (1 - r), \beta)$ where $r \geq p$

If the deontic constraints on my action are fully expressed by a Ramsey/Savage style model, then I have two options. Either, I

3. prefer $\alpha$ to $\beta$ and prefer $(r, \alpha; (1 - r), \beta)$ to $(p, \alpha; (1 - p), \beta)$

4. or I prefer $\beta$ to $\alpha$ and $(p, \alpha; (1 - p), \beta)$ to $(r, \alpha; (1 - r), \beta)$.

We can take the deontic operator $O(a)$ to mean that I ought to perform action $a$. If we take decision theory as presenting internalist deontic norms, then the idea is that if something is preferred then it is what I ought to choose. The above formulae seem, then, to suggest something like $O((3 \lor 4) \land \neg(3 \land 4))$. This is a coherence norm because it is about a certain kind of harmony between a group of preferences and doesn’t seem to say anything about instrumental norms, or what is a means to what. If decision theory presented instrumentalist norms, one would expect the utilities of the lotteries to be determined by the utilities of the subprospects $\alpha$ and $\beta$. But in every presentation of the expected utility theorem, that determination is symmetric—choices over lotteries equally determine the utility of the subprospects as choices over subprospects help determine the utility of the lotteries. (Dreier 2005) makes the same point:
In other words, a decision theorist computes the utilities of \([\alpha, \beta, \text{ and } \gamma]\) by inquiring into the agent’s preferences among \([\alpha \lor \beta, \beta \lor \gamma, \text{ and } \gamma \lor \alpha]\), and also her probabilities over \([\alpha, \beta, \text{ and } \gamma]\). The utility assignment that results is, as I noted in section 1, a theoretic construct that in effect summarized a lot of information about a coherent agent’s preferences and credences. So, far from being asymmetrically determined by my utility for \([\alpha, \beta, \text{ and } \gamma]\), my preferences over their disjunctions determine those utilities. At least, so it is according to the founding spirit of decision theory. (211)

Nevertheless, I think that too closely equating the deontic norms with the structure of models constructed in this manner is a mistake. Note first that neither Ramsey nor Savage, nor Luca and Raiffa were engaged in a project of presenting rational explanations of action according to our definition. They were not looking for the source of rules of behavior, or what those rules are from the first person point of view. They were only looking for a way to represent those rules with a model that was convenient and reasonably isomorphic with them. When we realize that the construction of utilities numbers is a different task than giving a rational explanation for why an agent ought to do something, then we can get a clear picture of how instrumentalist norms and decision theory might be compatible.

If we are giving a psychological reading of Ramsey/Savage style models, it behooves us to remember that, the way Ramsey puts it, he is looking for a way to measure degree of belief and value. While Savage is trying to define those terms, for him they are terms of art, and not necessarily associated with the concepts as the agent thinks about them, as we saw above. So with respect to Savage’s approach, insofar as it is similar to Ramsey’s, we can think of it as creating a system of measurement for degrees of belief and value as the agent understands them. So while it is the case that in order to measure the value of outcomes and the degrees of belief of an agent (without just asking him) we need to incorporate choices over both lotteries and sure-shots, that does not mean that the underlying concerns of actual practical rationality work that way, or that decision theory must be interpreted as implying that\(^8\). The underlying rational norms could have a structure that generates these patterns in preference in a way that is different from how the model reveals them.

---

\(^8\)Remember that, due to various complications, people like Ramsey, Savage, and Samuelson were trying to get away from psychological interpretations of decision theory, so defining degrees of belief and preferences in terms of behavior was an attractive alternative. Now that we are back to being comfortable with psychological interpretations of the model, there is less of a need to rely on behavior as the grounds for rational norms, or as our evidence of their existence.
One way this could happen is if there was not a wide scope norm governing (1) and (2), but rather a disjunction of narrow scope norms.

As a quick reminder, norms of coherence as such amount to the idea that the actual world is one where $O((3 \lor 4) \land \neg(3 \land 4))$ is true (you ought to do one or the other, you may do either, but you may not do both), and instrumentalism implies that it is one where $(O(3) \lor O(4)) \land \neg(O(3) \land O(4))$ is true (it’s either the case that you ought to do the one or it’s the case that you ought to do the other, and you may not do both).

(Dreier 1996) argues that if normativity is represented along the lines of the construction of utility functions in Ramsey/Savage models, the advice decision theory gives is something like “Attend to the axioms!”, and there are always multiple ways to bring oneself in line with the axioms. This means, that from the point of view of deliberation, an agent would find himself in a position where he reasons that given his desires, he could either decide to (3) or decide to (4) because the axioms say so, or that behaving in line with the axioms is coherent, or something like this.

The problem with this is that it is not at all clear why I ought to comply with the axioms. With Davidson and Lewis, there was a story about the linguistic community, or about the construction of my mental states, that was supposed to explain why all agents are under pressure to comply. If we take Dreier’s claim at face value, then it seems that we are left with the option where the axioms have intrinsic appeal from the first person point of view. But this is not at all clear. (Kolodny 2005) and (Kolodny 2007) provide some reasons to think that there is not much to recommend coherence norms from the stance of first-personal practical rationality. For my part, I would just like to note this: it is unclear why an agent couldn’t rationally disregard coherence thus defined. If we complain that an agent is unrepresentable in EUM and he says “So what?”, it’s not clear to me how the axioms are supposed to get any traction with his motivations—especially if he is careful to avoid Dutch books as they arise.

There are, however, some cases where it appears at first glance that an agent’s valuation of a lottery helps determine the value of the subprospects, as one would expect from Dreier’s approach (where ‘valuation’ here means actual psychological valuation, not just utility value). One might, for example, have a desire directly for the participation in a lottery. I might want to play poker, so my desires are not correctly described as solely over quantities of money. But even in such a case it

---

9 Another way of cashing out what it means to be coherent is to be invulnerable to Dutch books. But while it is true that if you obey the axioms you are invulnerable to Dutch books, the converse is unclear. It may be the case that an agent is invulnerable to Dutch books without obeying the axioms.

134
would not be appropriate to say that the lottery preference determines what value an agent places on the subprospects (possible poker winnings) by way of considerations in EUM. This is because if the participation in a lottery is something that an agent directly desires, it is an object of his desire, and, therefore, its own outcome. There may be a relationship between the desire to play poker and the values associated with different possible poker winnings, but it is not the relationship represented by the expected utility of the winnings. That expected utility is a more abstract idea. It represents the value of those winnings independently of the satisfaction of the desire to play poker. A real poker game, when a person actually wants to play, satisfies not only the desire for winnings, but the desire to play the game. The lesson here is that, on the instrumentalist picture, the strength of desire for subprospects, or, at the root level, outcomes, always determines the strength of desire for the abstract EUM lotteries in which they are embedded.

Furthermore, if we understand utilities as a system of measurement that reveals preferences that abide by instrumental norms, then we have a clear way that Ramsey/Savage models relate to rational norms—by assuming them and creating a measure that reveals details of the relationships between rational preferences. For example, suppose instrumentalism about rational norms is true—probabilistically coherent degrees of belief, strengths of desire and all. In this case, the agent’s preferences over the outcomes rationally determine his preferences for lotteries because probability is a measurement of the effectiveness at bringing about certain outcomes. This idea is reflected in (Mas-Colell, Whinston, and Green 1995), “The theoretical analysis to follow rest on a basic consequentialist premise: We assume that for any risky alternative, only the reduced lottery over final outcomes is relevant to the decision maker” (170). (Hampton 1994) also notes the consequentialist nature of the instrumentalist position. If we assume that this is true, then the structure of the Ramsey/Savage models make sense as methods of revealing strengths of preference and associating them with utility numbers. In order to create a numerical measurement of strength of preference, Ramsey/Savage models require observations of choices over lotteries in order to associate a distance with the strength of the preference for one outcome versus another. The reason that the theorist is justified in asserting that (3) or (4) holds is not that they are coherent, but that if the strength of the desire for $\alpha$ is greater than the strength of the desire for $\beta$ then a lottery with a higher probability on $\alpha$ will be more effective at satisfying the agent’s desires. When the theorist says that either (3) or (4) ought to be the case, she is not saying that the agent is in a situation where both (3) and (4) seem normatively appropriate to him; rather, the theorist is saying that given what she knows we are either in a situation where (3) is normatively appropriate from the agent’s point of view, or
(4) is normatively appropriate from the agent’s point of view, she just doesn’t know which without more choice data.

It seems to me that the founding spirit of decision theory was not one where first-personal psychological states were the intended topic—in fact, they were to be disregarded. But as we’ve seen, one cannot present rational explanations without talking about the agent’s own first-person experience. Once we have taken on board the idea that talking about the first-person is appropriate, we can see how models built in the Ramsey/Savage style determine utility numbers in a way that is completely different from how agents determine the expected value of lotteries and outcomes. Because of these considerations, it is a mistake to think that decision theory is more compatible with a coherentist rather than instrumentalist theory of normativity. Furthermore, instrumentalism has intuitive benefits over coherentism, like straightforwardly meeting the ‘So What?’ objection.

5.5 Armendt: Being ‘Of Two Minds’

Dutch Book arguments demonstrate that an agent’s preferences put them at risk of a sure loss if a clever bookie presents them with a group of trades that exploit their preference ordering. Generally, authors take exploitability of this type to be incompatible with being ideally rational, but there are different possible explanations for why rationality implies that we shouldn’t be exploitable. Armendt argues that the problem with exploitability is that it implies that the “the agent displays pragmatic inconsistency, in that he gives conflicting evaluations to the same option(s) [in different contexts]” (4).

A ‘conflicting evaluation’ of an option can take multiple forms, all sharing the similarity that an outcome contributes differently to the choiceworthiness of action in different contexts—a judgment that is made independently of an agent thinking that there is something about the context that warrants such a difference. The agent cannot be, for example, sensitive to global properties of gambles, and, as such, explain the differing evaluations in terms of how global properties alter the value contribution of outcomes to lotteries. Armendt wants to give an explanation of what is wrong with agents who violate the norms of EUM without good reason.

As long as the agent is not sensitive to global properties, a divided mind could be about degrees of belief—an agent might have probabilistically incoherent beliefs about a set of events even though he abides by the value aggregation rules of EUM (he always multiplies the probability of the event by the value of the subprospect to get the value of its contribution). He could have a divided mind about the value of the outcome, effectively giving it a higher utility in some gambles than in others.
Or he could have a divided mind about the contribution that an outcome at a given probability makes to the value of a gamble.

Dutch Books can be constructed to take advantage of an agent with any of the versions of a divided mind above, but I’ll just give one example using a divided mind about the value of an outcome to illustrate the point. Suppose, when making a choice between \( \alpha \) and \( \beta \), \( C(\alpha, \beta) = \alpha \), and furthermore, that this preference is strict, such that our agent would prefer \( \alpha \) over \( \beta + $1 \). Then, also suppose that our agent has preferences consistent with \( C(\beta, \gamma) = \beta \) and \( C(\gamma, \alpha) = \gamma \). This implies the following intransitive preference ordering: \( \alpha \succ \beta \succsim \gamma \succsim \alpha \). A clever bookie could offer this agent a group of trades that would guarantee the bookie his original position plus $1 from our exploited intransitive agent. Specifically, the bookie would offer the agent these trades in succession—\( \alpha \) for \( \beta + $1 \), \( \gamma \) for \( \alpha \), and \( \beta \) for \( \gamma \)—or he could offer the agent those trades as a group and then they could square up afterwards. In a case where the bookie starts with \( \alpha \), he ends up with \( \alpha + $1 \). And while the agent started with \( \beta + $1 \), he now has just \( \beta \).

Armendt’s thought is that the fundamental problem driving why the agent is not entirely as he ought to be is not that he is at risk of losing $1, but that our agent is of two minds about the value of \( \alpha \) in different contexts. Being at risk of losing money is just a symptom of his underlying disease. When he thinks about the value of \( \alpha \) relative to \( \beta \) or \( \beta + $1 \), he thinks of \( \alpha \) as more valuable than those options. In other words, his desire for \( \alpha \) generates ‘ought’ statements that imply a judgment that he ought to choose \( \alpha \) over \( \beta \) or \( \beta + $1 \) (or things of equivalent value). But when \( \alpha \) is compared with \( \gamma \), he judges that he ought to choose \( \gamma \), implying that the value or desirability of \( \alpha \) is lower than it was judged to be in the previous comparison.

Armendt says this is inconsistent and incoherent. With our additional framework from the previous chapter, we can say, more precisely, that this implies some kind of contradiction in this agent’s practical stance—the set of ‘ought’ judgments that he is inclined to make in virtue of his desires.

Being of two minds about something is supposed to be a violation of a norm of coherence as such. Armendt allows that there could be other types of norms working in concert with coherence norms, but divided-mind inconsistency is not reducible to them: “It is difficult to find principles more fundamental, judgments more secure, from which to argue that divided mind inconsistency is a flaw. Notice that the norm of consistency to which we appeal, consistency in evaluating our options, is not the pragmatic norm recommending that one always act so as to promote one’s goals. The norm that divided-mind inconsistency is to be avoided stands on its own, not on a justification by the pragmatic norm. It is instead a norm that regulates how we should conceive of, or specify, our interests (i.e. consistently)...” (5).
What’s great about this argument is that it does actually seem to support the right kind of pressure for an alternative to instrumentalism. It regards first-person deliberative considerations and their conceptually necessary components. After all, what can it mean to endorse a judgment that \( \alpha \) is of a particular value, and yet simultaneously endorse that it is another value? Once again, we can take two readings of the possible appeal of coherence. The first is that coherence is a necessary condition of the possibility of having the mental state at all, which seems, in this case, a straightforward interpretation. Alternatively, we could think of coherence as an object of desire—as directly appealing from the first person point of view. In such a case, the agent, upon recognizing that his judgments regarding the value of \( \alpha \) were inconsistent, would change them simply because they are inconsistent. But are either of these plausible options?

5.6 Three Problems with Norms of Coherence as Such

The best way to see the problem with Armendt’s argument is to examine the impossibility of his conclusion. There are ultimately three problems with norms of coherence as such, regardless of the route by which they are justified. They are as follows: these norms are not very appealing as aims or objects of desire, they are not necessary to having mental states in any ‘constitutive parts’ sense, and finally, the limited degree of freedom agents have in responding to relevant cases suggests that they actually face a disjunction of narrow scope norms. This means that Armendt’s position does not provide for a norm of coherence as such, but actually seems plausible because it is a slightly inaccurate description of a set of real narrow scope norms. More specifically, he gives us no reason to believe that the sense in which we ought to have coherent beliefs is anything above and beyond the sense in which our beliefs must be responsive to evidence, or the reasons for them. Evidential norms, I will claim, are narrow scope.

With respect to thinking that consistency is a necessary condition on having these mental states at all, note that it is perfectly possible to have inconsistent beliefs. We have them all the time. This also goes for inconsistent valuations of outcomes. The fact that Armendt’s position relates to judgments that agents make when considering different gambles makes this all the more plausible. If I am assessing a choice between \( \alpha \) and \( \gamma \), certain considerations may become salient to me that I tend to neglect when comparing \( \alpha \) and \( \beta \). This doesn’t mean that I cease to have beliefs, just that my mind is, in a very natural and common way, compartmentalized. Compartmentalization,
or the fact that I am capable of endorsing different beliefs, desires, etc., in different contexts where they become salient to me, is part of the explanation for why having inconsistent mental states is possible—those states, for one reason or other, just do not seem to interact with one another they way they are supposed to. The real question is, what do I do when I discover that I am compartmentalizing my judgments in this way? When these judgments do finally interact, what ought I do and why ought I to do it?

As always, we have a couple options for explaining the presence of a norm of coherence as such. Firstly, consistency might not be constitutive of mental states simpliciter, but it might at least be constitutive of them when they interact with one another and this compartmentalization breaks down. Secondly, it might be that consistency is independently attractive (or inconsistency unattractive).

Let’s take, as an example, the Preface Paradox. The example runs that a professor does a lot of research and writes a textbook containing a long list of facts about sea turtles. In the preface to his book, he writes that despite careful research, the scope of the book suggests that he must have made a mistake somewhere. This means that the professor has contradictory beliefs about his book. Individually, he believes all of the claims in the book. These claims, together, logically imply their conjunction. However, in the preface he articulates a belief on the basis of good evidence that the conjunction of the claims in the book must be false. So our professor has logically contradictory beliefs.

Insofar as we are comfortable talking in terms of graded beliefs, there is a common solution to this problem. The professor has a degree of belief in the conjunction of the claims in the book that is naturally lower than his degree of belief in each of the claims taken independently—the likelihood associated with the conjunction being the product of the likelihood of each of the independent claims. This seems to be the sense in which the contradiction between the preface and the content of the book maps onto a rational deliberative process—rationality requires probabilistically coherent degrees of belief, and in discussing the professor’s ‘contradictory’ beliefs, I was smuggling in a conception of binary belief, perhaps something like belief passing a threshold of certainty, or something like this. In this case, the contradiction is only in that the world cannot simultaneously satisfy all the claims in the book, not so much in the mindset of the professor, who is making a kind of limited assertion of these claims. The question for us is whether a norm of coherence as such plays a role in motivating the relationship between these contradictory claims and the degrees of belief that the professor adopts.

This type of analysis is not just specific to the Preface Paradox. As long as we endorse Bayesian conditionalization as an appropriate model for how to respond to
contradictory judgments, we will modulate our degree of belief in conflicting judgments on the basis of our assessments of their independent likelihoods. But this is, in this context, a narrow scope norm. I take beliefs as representatives of the evidence for them. If we take $B(A)$ to mean ‘believes $A$ above some threshold likelihood’, then my the appropriate response to contradictory beliefs is not the wide-scope $O(B(A) \lor B(\neg A) \land \neg B(A \land \neg A))$. Rather, the evidential story would suggest that I am either in a position where I $O(B(A))$ or $O(B(\neg A))$ on the basis of the evidence, or antecedent likelihood, of $A$ and $\neg A$. There seems to be no norm of coherence as such here that provides normative force additional to appropriately described narrow scope norms. This is a case of translatability, suggesting that we don’t have a need to be coherent as such. It just isn’t clear that probabilistic coherence is a result of a need to be coherent as opposed to, or independent of, a need to appropriately respond to evidence.

Furthermore, Armendt should agree with this. He thinks that Bayesian conditionalization is the appropriate way to update beliefs on the basis of new evidence—this is how we get the probabilistically coherent beliefs that he thinks we are under pressure to have. Any additional pressure to change my beliefs would deviate from the recommendations of Bayesianism. But, even worse for Armendt, as long as we maintain our standards for how efficacy effects value contributions, additional pressure to change my beliefs leads to exploitability and divided mind inconsistency.

For example, suppose that in addition to Bayesian considerations, I ought to reduce my degree of belief in both $A$ and $\neg A$ just because they contradict. So rather than $p(A) + p(\neg A) = 1$, they sum to less than 1. This could be considered an instance of coherence itself having an impact on my beliefs additional to the impact of narrow scope evidential norms. But then I can still be Dutch Booked! I would be willing to sell $\alpha$ at $[p(A) + p(\neg A)]\alpha$ and buy $\alpha$ for full price. But $\alpha$ is the same thing regardless of whether you get it straightforwardly, or from a lottery as the outcome of $A$ or $\neg A$. So this betting behavior would value $\alpha$ based on the context of its presentation, which is just the problem Armendt was trying to avoid in the first place. Any substantive difference here in the recommendations of norms of coherence as such and narrow scope evidential norms creates divided-mind inconsistency.\(^{10}\)

One might object that what counts as evidence for what is itself not a framework independent question. That the evidence we get can influence what we believe directly, or it can motivate us to change the framework that we are using to assess

\[^{10}\]This example assumes that other parts of EUM still hold for this agent, for example, how he aggregates the value of options in a lottery. But Armendt’s own examples work in the same way. I will not investigate here whether there is a way in which a divided mind of one kind can make up for a divided mind of another kind.
what it counts as evidence for. In such a case, we might think there is reason to say that coherence really does explain things at a foundational level inexplicable by narrow scope norms.

Davidson gives an example of walking into his neighbor’s house.

“Let me take another example: one drawn from real life, or at least from my life. One late Spring afternoon I was returning home from my work at Princeton University. It was a warm day, doors stood open. I lived in one of a row of attached houses in which faculty members were housed. I walked in the door. I was not surprised to find my neighbor’s wife in the house: she and my wife often visited. But I was slightly startled when, as I settled into a chair, she offered me a drink. Nevertheless, I accepted with gratitude. While she was in the kitchen making the drink I noticed that the furniture had been rearranged, something my wife did from time to time. And then I realized the furniture had not only been rearranged, but much of it was new—or new to me. Real insight began when it slowly came to me that the room I was in, though identical in size and shape to the room I was familiar with, was a mirror-image of that room; stairs and fireplace had switched sides, as had the door to the kitchen. I had walked into the house next to mine” (Davidson, “Incoherence and Irrationality”, p. 191).

The two judgments, that he is in his own house, and that he is in his neighbor’s house, contradict. Furthermore, what Davidson saw he took as evidence for the claim that he was in his own house and that the furniture had been rearranged. Ultimately, a body of evidence that had been conceptualized as suggesting that he was in his own house was then reconceptualized as suggesting that he was in a different house. In such a case, it might seem that rationality presented Davidson with a wide scope doxastic norm—either believe that you are in your own house with the furniture rearranged, or readjust your framework for understanding the evidence and judge that you are in your neighbor’s house. If it seems plausible that this situation allows for the kind of freedom inherent in a wide-scope norm, then we have reason to believe there is a norm of coherence as such operating here.

In response, first notice that if he had judged that he was in his own house with the furniture rearranged, he still wouldn’t be entirely as he ought to be—specifically, he would have false beliefs about what house he was in. So coherence is not, in any obvious way, all there is to understanding the doxastic normative pressure on Davidson in this situation. In such a case, this lack of freedom in determining his
response suggests that he is facing a disjunction of narrow scope norms rather than a wide scope norm.

One might object at this point that correctness is just too stringent an epistemic standard. Surely we can rationally update our beliefs and still, through bad luck, end up wrong.

This seems correct. So it is worth noting what would be entailed by a wide-scope norm here. If Davidson was governed by a wide-scope norm in this case, then that should also be true in similarly structured cases. Indeed, almost every situation I find myself in is an opportunity to reevaluate what my evidence is evidence for. For example, I might take the fact that I see a laptop in front of me not as evidence for there actually being a laptop, but, through reconceptualizing it, as evidence for the idea that I am hallucinating, or being deceived, or any number of such things. And yet, that kind of freedom to reconceptualize my evidence seems inappropriate. There seem to be narrow scope norms constraining when I am permitted to take my evidence as something other than face value.

But this is just a piece of the puzzle. Our question is about whether coherence itself played any role here—whether in appreciating the pressure to either revise the framework for understanding the evidence, or believe he was in his own house with the furniture rearranged, he was responding to a norm of coherence as such that was in some way independent of a narrow scope evidential norm.

I will not attempt to give a full account of what goes into norms for determining how an agent ought to respond to evidence, or form frameworks on the basis of evidence. But I do want to demonstrate that, at least at first glance, there seems to be no additional role to play for coherence beyond a narrow scope description of the normative pressure Davidson was under. Taking $H$: “I’m in my house with the furniture rearranged”, and $N$: “I’m in my neighbor’s house”, rather than the evidence presenting Davidson with a wide scope norm $O(B(H) \lor B(N) \land \neg(B(H \land N))$, he might face a disjunction of narrow scope norms. Either he is under pressure to believe that he is in his house, or he is under pressure to believe he is in his neighbor’s house, he just doesn’t know which—$O(B(H)) \lor O(B(N))$. This disjunction of narrow scope requirements seems to be a natural extension of the narrow scope demand to believe what is true. I know that normative pressure pushes me in one direction or the other, I just don’t know which way. Any plausible coherence norm in this arena could be explained by this narrow scope norm in conjunction with ideas about how I cope with my own epistemic limitations, and not vice versa. This would imply that the coherence norm was not a norm of coherence as such because it is not the most fundamental explanation of what is going on. And that’s all there is to the story. We do not need an additional norm of coherence as such to describe the pressure.
that Davidson is under, and these narrow scope norms seem perfectly plausible and attractive from the first-person point of view.

When Armendt says that there is no other judgment as obvious as ‘that divided-mind inconsistency is bad’ he overlooks some important options—that you should be effective at doing what you’re aiming to do, whether that is having accurate beliefs or satisfying your desires. The badness of divided mind inconsistency is then explicable in terms of instrumental norms. If an agent comes to different conclusions about what is most effective in two different contexts, then he must have made a mistake in one instance or the other.

5.6.1 Radical Inconsistency

The Preface Paradox and Davidson’s neighbor’s house are two cases where norms of coherence as such seem implausible. But advocates of norms of coherence as such who think that they are essential to the constitution of mental states are mostly concerned about cases of extreme incoherence. Davidson and Preface Paradox are not cases where an agent would be at risk of no longer having beliefs and desires at all. So it might be that those cases are explicable by instrumental norms whereas more severe cases are explained by norms of coherence as such.

The problem with a view like this is that these additional norms of coherence as such are just unnecessary. The Teleological Functionalist position I am advocating for suggests that an essential part of having these mental states is for them to create these narrow scope normative pressures. So no matter how incoherent an agent becomes, as long as he has beliefs and desires he will be subject to narrow scope deontic norms that recommend something that will fix that incoherence.

5.7 Instrumentalism, Parity, and Exploitability

Given that I have rejected the possibility that there are any norms of coherence as such, the problem still remains as to why agents should be as precise in their choices as EUM seems to suggest that they should be. My answer to this is that they don’t actually need to be. Because I have to perform actions that are maximally precise, and my strengths of desire are not that precise, utility functions are just approximations of the vague demands of instrumental rationality. Many similar utility functions will be appropriate given the strengths of desire that I actually have.

That strengths of desire are vague seems self evident. The problem is that it creates situations where certain choices are not decidedly better, worse, or equivalent
to, other options. It might be difficult, for example, to tell whether the distance between my desire for apples and my desire for pears is greater than that between pears and bananas. As such, it would be difficult, just by reflecting on my strengths of desire, to be consistent in my choice of lotteries between them—or to be consistent in the ‘ought’ judgments I make as a result of them. The vagueness of a strength of desire is similar to the vagueness in the assessment of how sweet something tastes. In each case, I face a problem of translating an experience of what something is like—the first-personal experience of the appeal of that outcome, or what it is like to taste something—into some kind of linguistic representation. This is what happens when I translate my strengths of desire into a ‘to-a-degree-U’ judgments with the \( v(\phi) \) function, or when I verbally describe how sweet something tastes to me.

(Chang 2002) investigates these types of cases and concludes not that strengths of desire are vague, but that normative pressure is not matter of there being more, less, or equal pressure to take two alternative courses of action. She argues that there is a fourth type of comparative relation that isn’t reducible to the other three—that two options can be on a par, meaning that it is permissible to choose one or the other even though they are not strictly speaking equal. One example of a slight sweetening is that a coffee drinker might prefer sweeter coffee to that which is less sweet, but be indifferent between adjacent cups of coffee that are ± one grain of sugar. If he is indifferent between a change of one grain of sugar, then by the transitivity of preference he ought to be indifferent between all levels of sugar in coffee, but he is not. Another case is where a young man is choosing between a career in the military and a career as a priest. He has so much trouble deciding between them that he figures that he could go either way. But suppose the general stops by his house and offers him a chocolate chip cookie to join the military. The military career plus the cookie is now better at satisfying the young man’s desires than the old military career. And if the old military career was as good as the career in the church, then the cookie should tip the scales in favor of joining the armed forces, but that seems odd.

I would like to note two objections to the possibility of parity before moving on with the hypothesis that strengths of desire are vague and that this vagueness can contribute to explanations of the above cases. Vagueness of strengths of desire falls into a broader category that I will call resolution problems. An agent faces a resolution problem when a decision is put to him in a way that goes beyond his ability to consistently distinguish between the alternatives in terms of how well they satisfy his desires.

Resolution problems are not a problem for the notion of instrumental rationality. Resolution problems are just problems that rational agents face in determining what
they ought to do—just like if a situation required complicated math, or seeing in a dark room. In the sugar case, the agent will likely take a sip of a cup of coffee at one point and realize that the coffee is getting better. His inconsistency in noticing it getting better is a problem of his tongue, not of his rational deliberation\textsuperscript{11}. In the military case, the young man’s desires probably present their demands to him in terms that make it difficult to assess which career will best satisfy his desires—as in terms of fulfillment, justice, happiness, etc. There is so much uncertainty in his ability to map those properties onto his career prospects that the cookie could easily fail to increase the value of the military career enough to cover the difference between it and the church. If we decrease the size of the uncertainty, like as between a cup of coffee and a cup of tea, then the cookie can easily tip the scales\textsuperscript{12}. Resolution problems do not reveal that normative pressure admits of parity, but that in some cases it is difficult to satisfy the demands of normative pressures due to lack of information or sensory resolution.

One final reason to think that resolution problems, rather than parity, explain these cases, is that parity suspiciously only occurs in cases of very close value comparison—cases where you would expect resolution problems to arise. There are no plausible cases of parity arising between $5$ million and a fresh apple, for example. But if parity is a fourth kind of relation, independent of the expression of the other three, then there is no obvious reason to think that this would have to be the case. Being worse than, better than, or equal to, are functionally interrelated in a way that leaves no space for another relation. This fourth relation, if it existed, would seem to exist somehow outside of the functional definition of the other three in terms of one another, and would apply in cases where the application of the others was inappropriate. This does not leave us with an understanding of why parity would only arise in cases close to the appropriate application of equality.

Leaving parity, we can adopt the idea that strengths of desire are vague. But we still have our remaining problem that agents risk being exploitable as a consequence of that kind of imprecision.

\textsuperscript{11}Another way of phrasing this that I would be perfectly happy with is to say that you inconsistently taste it as sweeter. So it is not that there is betterness independent of the taste that you are inconsistent at detecting. Rather, it is that you have a better experience inconsistently with respect to the stimulus that causes it.

\textsuperscript{12}Of course, there are certain ways of dealing with the uncertainty involved in the career case that imply that our young man must be violating expected utility theory. Without committing one way or the other, I would like to note that one way around this would be to say that the young man does not have enough information to make a probability assignment to the proposition that one career is better than the other, and as such is not in a position to make an expected utility calculation.
In actuality, vagueness doesn’t imply that an agent will be exploitable to any serious degree. Each individual cycle of a Dutch Book would leave an agent “worse off” within the vague boundaries of his strengths of desire—so, not necessarily worse off from his own point of view at all. But multiple cycles of books like this would risk leaving the agent worse off from his own point of view—for example, by getting him to trade away all the sugar from his coffee.

Some authors address exploitability by suggesting that agents ought to evaluate all the choices in a series as one decision. This would mean that rather than evaluating the value of each instance of a Dutch Book individually, the agent would assess the choice to trade away a grain of sugar as part of a larger series of choices that will leave him worse off in the end, and that is why he shouldn’t participate in the first instance of the series. While this would avoid exploitability, it is not a viable strategy. This is because, as noted in previous chapters, to have a preference is to have an all things considered formal desire, which is to judge that something is better than the alternative. When an agent makes such a judgment, they are normatively constrained to follow through and act upon that judgment. So if the agent does actually have a preference to follow through with each instance of a Dutch Book, then they ought to do so. There is no perspective available to them at that time that can explain normative pressure to do otherwise. In short, the problem with this approach to avoiding exploitability is that it is not in any clear way compatible with the agent’s actual preferences and the structure of his reasoning process.

An agent could potentially see that a particular series of choices could be incorporated into a Dutch Book and on that basis decide that the series of choices was not a good one. But if that is the case, then the agent’s judgments about what is good have changed—either there is a standing desire not to be Dutch Booked that is outweighing, based upon its strength, the desire to act out each instance of the series, or the agent has ceased to endorse his desire for each instance. Otherwise, the agent wouldn’t see anything wrong with being exploited. In either case, this situation is appropriately represented as a preference change. Because there is a preference change, there is no possibility of a Dutch Book, and no need for the theoretical apparatus of evaluating a series of decisions holistically in opposition to preferences at a given time.

To reiterate, all of this reasoning is based upon the necessity of using the formal notion of desire, and a notion of preference based upon that, in order to pass the ‘So What?’ objection. If agents were able to ask further questions, beyond their preference ordering, regarding what they ought to do, then decision theory would not be a complete and autonomous rational explanation—the formal model would come with caveats about situations in which an agent shouldn’t do what he seems
to prefer in the model. But all of this can be avoided by using a formal notion of desire, and thinking of holistic decision making as more appropriately represented as a case of preference change.

But how do these notions of holistic decision making apply to our case of vague strengths of desire? In each instance, the agent risks viewing going through with the book as what he ought to do, and as such he would risk continuing down his path to no sugar. So from the point of view of all of the utility functions appropriate to him, and more importantly, from the point of view of maximizing the satisfaction of his vague desires, he is under pressure commensurate with expected value to behave in a way that more closely resembles one utility function. The utility function he chooses could be any of the utility functions that are appropriate to him. While each utility function would recommend adopting itself, his desires—the things actually creating normative pressure—are vague, and, as such, would recommend adopting any of the utility functions appropriate to him. Acting in accordance with any of those functions would more effectively satisfy his desires than being exploited by a Dutch bookie. In other words, it would be outside the boundary of the vagueness of his strengths of desire not to behave in a way that more precisely tracks one utility function.

But this only follows if he is actually at risk of being Dutch Booked. This is because the loss from the Dutch Book is only as bad as its expected value. If I don’t expect to be Dutch Booked, then there is no expected loss, and maximizing my expected desire satisfaction can safely happen within the vague boundaries of a set of utility functions. It is only when actual Dutch bookies, or some equivalent, are around that I risk failing to maximize my desire satisfaction by having a vague practical stance.

5.8   Conclusions

The role of this chapter was to evaluate and oppose alternative theories of rationality that employ norms of coherence as such. I have argued that there is good reason to think that these types of norms do not exist, because whether we talk about them as insinically attractive or as constitutive of the mental states they govern, they seem to be at best a redescription of a more fundamental instrumental norm. I then addressed the fact that instrumental norms leave some vagueness in the demands of instrumental rationality because desires are not maximally precise. I argued that this is not a large problem, either for this as a theory of rationality or for instrumentally rational agents, because it does not imply any substantive exploitability by Dutch Bookies.
Chapter 6

A Defense of the Independence Axiom

6.1 Introduction

As a quick reminder:

Independency - for any three lotteries \( l_1, l_2, l_3 \in L \), if \( l_1 \succ l_2 \) then for any \( p \in [0, 1) \), \( (1 - p, l_1; p, l_3) \succ (1 - p, l_2; p, l_3) \).

Independency is the most controversial of the EUM axioms. And while it follows from the theory I presented in chapters 1-3, one might be inclined to modify that theory to remove this feature. So in order to defend the view I presented, I want to examine what rationality without Independency would look like. In the end, I will argue that the costs of rejecting Independency outweigh the costs of keeping it.

In *Risk and Rationality*, Lara Buchak presents an alternative to Expected Utility Maximization, Risk-Weighted Expected Utility Theory (REU). REU rejects Independency. It is based upon an understanding of rational agents as evaluating “neighboring pairs of outcomes” and judging “how much that chance of obtaining the better outcome adds to the instrumental value of the gamble” (54). The key idea is that the agent has a large degree of freedom in determining how this incremental chance of doing better impacts overall value. On EUM, this incremental chance contributes its expected value to the overall value of a lottery. On REU, this contribution is modulated by a risk function that an agent is free to define for herself. Because a “chance of obtaining a better outcome” is relative to what other outcomes are in a lottery, REU violates Independency by allowing a \( p \) chance at \( \omega \) to contribute
different values to different lotteries depending upon the other possible outcomes.

Buchak notes that one way of getting a clear sense of what REU amounts to is to notice that there are three questions that a rational agent must answer in order to figure out the value of a lottery. This is in opposition to the two that EUM asks an agent to answer. The three questions are

1. “First, an individual determines which ends he wants, and how much: these are the values captured by the utility function.”

2. “Second, he determines how likely various actions are to lead to various ends: how effective each means would be of realizing each particular end. This evaluation is captured by the probability function.”

3. “Finally, he determines which strategy to take when choosing among actions whose outcomes are uncertain: how effective each means is towards his general aim of realizing an end he wants. Specifically, he evaluates the extent to which he is generally willing to accept a chance of something worse in exchange for a chance of something better” (37).

EUM only seems to require the agent to answer the first two questions, assuming that there is only one acceptable answer to the third.

The freedom the REU agent has in assessing how a chance at a better option impacts the overall value of a gamble is represented by his risk function, \( r(p) \). The risk function also helps to illustrate a way in which REU agents think in terms of incremental opportunity over the worst possible outcome. It factors into the REU utility function as follows:

\[
REU(l) = u(\omega_1) + r\left( \sum_{i=2}^{n} p_i(u(\omega_2) - u(\omega_1)) \right) + r\left( \sum_{i=3}^{n} p_i(u(\omega_3) - u(\omega_2)) \right) + ... \\
+ r(p_n)(u(\omega_n) - u(\omega_{n-1}))
\]

Expected Utility Maximization can be presented with a similar incremental structure.

\[
EU(l) = u(\omega_1) + \sum_{i=2}^{n} p_i(u(\omega_2) - u(\omega_1)) + \sum_{i=3}^{n} p_i(u(\omega_3) - u(\omega_2)) + ... \\
+ (p_n)(u(\omega_n) - u(\omega_{n-1}))
\]

The primary difference between EUM and REU at this level is that the risk function represents a freedom in rationality that is part of the REU interpretation.
This freedom to further define the way in which a probability of incremental gain is relevant to decision making is not allowable on EUM.

This raises questions beyond just the formal structure of the theory. Given the fact that both EUM and REU must be tied to psychological interpretations, we need to understand how the REU agent’s psychology generates this risk function. In order to fully grasp what is going on with the REU agent, we need to understand what exactly reasoning is like from his point of view, and what exactly it is that EUM can’t capture about the reasons for action as the REU agent sees them.

The initial motivation behind alternative theories like REU is that EUM doesn’t capture people’s intuitions about how they ought to behave in all cases. Allais’s Paradox is the classic example. When choosing between \( l_1 \) and \( l_2 \), agents seem to have preferences that are incompatible, under EUM, with the choice they make between \( l_3 \) and \( l_4 \). Not only do they feel inclined to make this choice, but they would claim that these choices are perfectly rational.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2-11</th>
<th>12-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l_1 )</td>
<td>$ 0</td>
<td>$ 5M</td>
<td>$ 0</td>
</tr>
<tr>
<td>( l_2 )</td>
<td>$ 1M</td>
<td>$ 1M</td>
<td>$ 0</td>
</tr>
</tbody>
</table>

Buchak suggests that Allais might find the following reasoning intuitive:

[in the choice between \( l_1 \) and \( l_2 \)] He reasons that the minimum he stands to walk away with is the same either way, and there’s not much difference in his chances of winning some money. So, since \( l_1 \) yields much higher winnings at only slightly lower odds, he decides he would rather have \( l_1 \). . .

[between \( l_3 \) and \( l_4 \)] He reasons that the minimum amount that he stands to win in \( l_4 \) is a great deal higher than the minimum amount he stands to win in \( l_3 \), and that although \( l_3 \) comes with the possibility of much higher winnings, this fact is not enough to offset the possibility of ending up with nothing. So he decides he would rather have \( l_4 \). (12).

These preferences seem to violate EUM. No matter what utility we assign to $0, $1M, and $5M, we cannot represent these preferences as a case of maximizing expected utility. Not only are these kinds of preferences unrepresentable by EUM, but Buchak thinks that they are a competent expression of instrumental rationality—that the quote above is a description of reasoning about what is most effective at satisfying Allais’s preferences.
Buchak’s ultimate endorsement of REU over EUM is based on recognizing certain tradeoffs in theory choice that she thinks tip the scales in favor of REU. One of the strange consequences of REU, which Buchak admits, is that it implies that in certain situations an REU agent will pay money to avoid free true information that is relevant to the outcome of their choices. Buchak takes the strangeness of this conclusion to be more than counterbalanced by the fact that EUM cannot capture sensitivity to global properties. This sensitivity allows for the “paradoxical” preferences in the Allais case.

Perhaps it is a somewhat unpalatable upshot that more information is not always better for decision making, and that we cannot always say which of two informational positions is better even when the information in one is a superset of the information in the other. But I think it is more palatable than the claim that rational agents can’t care at all about the structural properties of the acts they are deciding among, particularly when we notice that the incomparability of informational positions arises from instrumental reasons for preference rather than from the values of the outcomes themselves. So I am willing to accept this upshot. If the reader gets off the boat here, then I have at least made it clear what tradeoffs are involved in choosing between permitting global sensitivity and disallowing it: the bullet one has to bite in accepting global neutrality as the uniquely rational norm is to reject the reasons for caring about global properties discussed thus far, and the bullet one has to bite in accepting that global sensitivity is permissible is to reject the idea that choices are always better from a position of more information (210).\footnote{Another benefit of Buchak’s approach is that in providing a representation theorem for the REU agent, she elucidates potential roles for beliefs and desires in an alternative to EUM. But because EUM also has a representation theorem, the debate between these two positions is even on that point.}

We face this tradeoff, in part, because the traditional conceptual arguments that Independence links the quality of outcomes and the value of lotteries don’t quite work. Buchak points out that the approaches Broome and Samuelson take to arguing for this connection completely miss the possibility that motivates REU—that the weight of reasons can change depending upon the possible outcomes in a lottery.

Samuelson argues for the rationality of compliance with Independence by noting that what happens in one state of the world cannot ‘contaminate’, or impact the goodness of, what happens in another state of the world (Samuelson 1952, p. 672-673). So if l₁ is better than l₂, that relationship will still hold when they are parts
of lotteries involving the possibility of $l_3$. Because the outcomes of each lottery are not simultaneously realized, their goodness should be evaluated independently, and so we should abide by this axiom.

Broome argues that failures of Independence are incompatible with the reasons for making choices. The idea is that if the agent judges there to be reason to pick $l_1$ over $l_2$ (or that the net strength of reasons favors $l_1$ over $l_2$), then those reasons will still net in favor of $(1 - p, l_1; p, l_3)$ over $(1 - p, l_2; p, l_3)$, since all reasons pertaining to $(p, l_3)$ will cancel each other out (Broome 1991, p. 95-96).

Buchak agrees with these points insofar as $l_1$, $l_2$, and $l_3$ are sure-shots at single outcomes. But in cases where they are lotteries with multiple options, REU allows for deviation from Independence. $l_3$ might, for example, introduce a new worst case outcome that changes the way in which the agent evaluates the relative contributions of $l_1$ and $l_2$ to the value of the overall compound lottery.

Samuelson fails to appreciate that there is a possible conceptual distinction between how good something is when it obtains and the weight it puts on an agent to choose a lottery at any particular probability. Broome fails to recognize that global properties can either give an agent new reasons, or change the way in which a given set of reasons weigh against one another. In either case, they may have found these assumptions plausible, but they did not explicitly argue for them, which is what would be necessary to combat REU.

The difference between what I am doing here and what Samuelson and Broome did is that I want to argue that the assumptions that Broome and Samuelson made, implicitly and without defense, are correct. We can see this by looking at what Independence violation implies about activities on the periphery of rational choice—things like representing the decisions of a group, justifying your decisions to others, and executing long term plans. There are also common practices in our culture that assume people obey Independence and that are not publicly criticized on that basis. One might think that if there were many REU agents out there, practices like this would be met with REU-specific criticism. The counter-intuitive results of Independence violation ripple out to the extent that the REU agent begins to look alien.

### 6.1.1 Thesis and the Plan

Regardless of whether one takes Buchak’s considerations to weigh more in favor of REU or EUM, I will argue that the correct way to view the tradeoffs between EUM and REU tips the scales further in favor of EUM. Firstly, I will argue that EUM can in fact capture a form of sensitivity to global properties, and it can do so in a way that
captures the most convincing cases. Furthermore, in arguing for the irrationality of deviating from EUM, I will implicitly defend the idea that the kind of global property sensitivity that EUM can capture is the only type we should want to capture. In this way, the apparent costs of accepting EUM as a theory of rationality disappear.

As I mentioned above, I will argue for the irrationality of Independence violation by looking at the impact sanctioned Independence violation would have on concepts in the periphery of decision making. I’ll take this approach because, as evidenced by the existence of this debate, people’s intuitions about what seems like a good thing to do in a particular situation vary widely. My contention is that this could be because irrational psychological tendencies might consistently warp our perception of what we ought to do. But if we look at concepts and practices related to decision making we might be able to avoid this influence.

I will look at four categories of examples that suggest human agents broadly aspire to obey Independence: how we reason about the will of a group, how we make inferences about efficacy, how we evaluate and execute long term plans, and how people respond to real world common practices. All of these examples create strange implications for Independence violators themselves, or suggest that people who think they ought to violate Independence don’t make up a large portion of the population.

REU is not the only way of rejecting independence. Other theorists raise objections to it as well. But REU is a good example that can help us get clear about what rejecting Independence might look like. REU is also relatively conservative for a theory that recommends Independence violation—recommending it only in some situations involving compound lotteries. So arguments that work against REU will tend to work against other kinds of Independence violators. Furthermore, the objections I raise follow mostly from the rejection of Independence itself, and not from the particularities of REU. But I won’t go so far as to say that these arguments work as a completely general defense of Independence because while the REU theorist doesn’t himself have the resources to avoid the situations I bring up, it is an open possibility that another way of violating Independence might bring solutions to the challenges I raise. The primary difficulty for a theory that attempts to do this would be to meet my objections in a way that is not ad hoc, and somehow flows from a reasoning process that could accurately be described as violating Independence. I’m not at all sure this is possible, so I will content myself to use the REU agent as our benchmark Independence violator.

---

2See (McClenen 1990)

3One challenging alternative to REU is a context sensitive Independence violator. One who might not violate Independence in the cases I bring up later on, but who might violate Independence in other cases. I do not take myself to present a complete argument against this position, but I will
It might be easy to explain why agents fail to conform to Independence in practice. They might make any number of mistakes along the chain of practical reasoning. But endorsing Independence also raises the more difficult question of why so many informed theorists reject it if it is in fact a real constraint of rationality. Even on an interpretation of EUM, there must be something about practical reasoning that can explain why Independence is so controversial. As it turns out, Teleological Functionalism just makes this question harder to answer. The most natural solution—that there is an irrational decision making system operating alongside the rational one—won’t work given TF. At the end of this chapter I will offer some possible explanations for the controversy surrounding Independence that are compatible with TF.

This chapter is broken down into eight additional sections. The first investigates the extent to which EUM can capture sensitivity to global properties. The second gives a TF interpretation of REU agents, contrasting their reasoning process with EUM and explaining the sense in which they can achieve a greater sensitivity to global properties. The next four sections investigate examples that suggest human rationality includes Independence. I then conclude with considerations governing an error theory for reflective Independence violators and a section on general conclusions from the project as a whole.

6.2 EUM and Global Property Sensitivity

The reason it is worthwhile to capture certain features of global property sensitivity is not because a theory of rationality needs to explain cases like the Allais preferences as such. Rather, it is because there are cases that are much less controversial that one might want to capture. As long as an agent can have any desire at all, then Allais preferences can be understood as EUM preferences by redifferentiating outcomes to include global properties that agents might care about. The problem with this approach is that there are no obvious candidate global properties that make this strategy plausible. In this section I want to highlight the uncontroversial cases that are easy to explain using EUM desires for global properties, and then contrast those cases with the controversial cases that can’t be easily explained away using this method (like the Allais Paradox). At the end of this chapter, I will offer several options for why Allais preferences might arise even though the traditional EUM approaches in this section cannot explain them.

The least controversial motivations for including global properties in outcomes come from sensitivity to particular histories. For example, if an agent wants to raise additional concerns that a position like this must overcome.
become president by being fairly elected, versus by rigging an election, or if a father can only save one of his children, but doesn’t think it is appropriate to flip a coin to choose which one. In each of these cases it seems perfectly possible that a rational agent could care about these things. But it also seems perfectly possible that this is just one of the desires that the agent has. He wants to become president by way of being justly elected. Or, he wants to save one of his children, and, if possible, he wants to do so by means of a process that respects the gravity of the situation. We can explain the difference in the agent’s behavior by making these different outcomes based upon the desires which would or would not be satisfied by taking a particular action—e.g., being justly elected is one outcome, being unjustly elected is another. So there seems to be no problem here, either for EUM or my interpretation of it.

Chapter 2 concluded in agnosticism about whether desires were a normative response to something else. Maybe basic desires need to be a response to desirable internal states—things like happiness, physical pleasure, health, etc. On the other hand, maybe desires can be for anything—a saucer of mud, or a twig of mountain ash, for example (Anscombe 1957). This latter case allows for a great deal of global property sensitivity. If someone most desires choosing lotteries that involve only three members, for example, then practical rationality is about describing his most effective means to satisfying that desire—a desire that is specifically about a global property.

On the other hand, if basic desires must be responses to desirable internal states, then this restricts the kinds of global property sensitivity EUM allows. Suppose, for example, that someone had a desire to avoid regret. Regret is something that might track global properties. If I pass up a sure shot at $1 million for a risky shot at $5 million, that might be an opportunity for regret.

This suggests that there may be EUM versions of Allais preferences depending on what is going on in the agent’s head. If an agent would feel terrible if he passed up a sure shot at $1 million, that could motivate a choice of \( l_4 \) and not stand as a corresponding motivation for \( l_2 \). Because avoiding regret is something the agent wants, it is an object of his desire and we should differentiate outcomes to include that fact. In a case like this, Allais preferences don’t violate Independence because the outcomes are too different. Now, \( l_3 \) includes an opportunity for regret.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2-11</th>
<th>12-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l_3 )</td>
<td>$ 0 + Regret</td>
<td>$ 5M</td>
<td>$ 1M</td>
</tr>
<tr>
<td>( l_4 )</td>
<td>$ 1M</td>
<td>$ 1M</td>
<td>$ 1M</td>
</tr>
</tbody>
</table>

Most theorists would agree that EUM can capture a degree of global property sensitivity. The uncontroversial cases I mentioned above are cases that EUM should
be able to capture through the redifferentiation strategy. The problem is that there are reasons to think that the regret based explanation of the Allais paradox doesn’t work, and there are additional cases that EUM can’t capture.

Firstly, notice that the differentiating outcomes method of rationalizing risk sensitivity would require the expected utility of passing up a sure shot at $1 million to be strangely high. For an expected utility maximizer to choose $l_3$ over $l_4$ because of regret, regret would need to completely offset the utility associated with a ten times better chance of $5 million over $1 million in the tickets 2-11 case. Even if regret is unpleasant, this seems like too much (Buchak 141).

With respect to other examples, suppose Jeff faces the following decision problem subject to two coin flips.

<table>
<thead>
<tr>
<th>HH</th>
<th>HT</th>
<th>TH</th>
<th>TT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deal 1</td>
<td>Elvis Stamp</td>
<td>Elvis Stamp and Gloves</td>
<td>Nothing</td>
</tr>
<tr>
<td>Deal 2</td>
<td>Elvis Stamp</td>
<td>Elvis Stamp</td>
<td>Gloves</td>
</tr>
</tbody>
</table>

If Jeff prefers deal 2 to deal 1 (assuming the goods are valued independently and have no complimentarity), then our redifferentiation strategy seems to imply that either Jeff adds a premium for the certainty he gets in the deal 2 outcomes, or he subtracts a tax for the regret he would feel in the deal 1 TH state. But Buchak argues that this is not the only possible way Jeff might take the situation. He might prefer deal 2 in a way that does not imply anything about a change to the value of the outcomes. The evidence for this is that Jeff might prefer deal 2 even though he would not pay money to get “nothing when he could have had gloves” over “nothing when he wouldn’t otherwise have had gloves” (if that were possible) (134). So Jeff does not endorse the comparison that is implied by the redifferentiation of outcomes, even though he still prefers deal 2. The suggestion is, then, that there must be something else going on.

There is another reason why the kind of risk sensitivity that an REU agent has is not fully capturable by EUM. In general terms, there is no such thing as the fixed contribution from an outcome to the value of a lottery on REU. If I were to try to construct a set of redifferentiated outcomes such that for each new outcome, it contributed the same value to the total value of a lottery as the original outcome in REU, it would be, in some cases, impossible. This is not entirely surprising because outcomes only have a fixed contribution to the value of a lottery in REU so long as the ranking of actions by event does not change. Because the risk function is sensitive to these rank orderings, the way in which the presence of an outcome contributes to the value of a gamble on REU is more complex, so there isn’t a straightforward
comparison here (148)\textsuperscript{4}.

\section*{6.3 Basic ‘Ought’ Judgments in REU}

The differences in reasoning and recommended outcomes between REU and EUM agents raise the question of what a Teleological Functionalist interpretation of REU theory would look like. It turns out that we need to make two changes to the EUM picture.

Firstly, Buchak argues that the risk function is not itself an expression of beliefs or desires. On REU, the utility function represents desires, probabilities represent degrees of belief, and the risk function corresponds to an entirely different kind of attitude directed at how the agent is inclined to structure the potential realization of his aims (56). This means that the first change we would need to make to the TF picture from chapter 2 is to include another mental state that is responsible for an agent’s risk function.

Secondly, if we keep the component of the chapter 2 interpretation where desires recommend judgments with ‘ought’ content, and that ‘ought’ content logically implies something about my preferences, then on REU that content needs to be more restricted. This is because the REU agent can’t reason to an action from desires and degrees of belief alone. He needs to ask a third question in order to set his risk function. So on a TF interpretation of REU, the content of basic ‘ought’ judgments from desires and degrees of belief needs to be insufficient for an agent to logically deduce conclusions about what he ought to do. An REU agent can only determine what to do once considerations from this third mental state category are brought in as premises for practical reasoning. Because the content of degrees of belief is fairly straightforward, the relevant modification is to restrict the content of the basic ‘ought’ judgments REU desires incline agents to form.

Without the first modification, the agent would have too much freedom in his decisions over lotteries to be REU. Without the second modification the agent would either reach conflicting judgments about what to do, or his new REU mental state from the first modification would be irrelevant.

On REU, utilities are determined by evaluating an agent’s preferences over lotteries that hold fixed the ranking of actions given the state of the world. For example,

\footnotesize{\textsuperscript{4}In section 8 I offer a possible error theory for REU reasoning that captures the way Jeff wouldn’t pay anything to get “nothing when he could have had gloves”. This is by specifying Jeff’s desires in terms of lottery profiles that are attractive to him. In cases like this, REU agents are willing to pay to satisfy their desire for particular global properties.}
bringing an umbrella and not bringing an umbrella change rankings when it is raining outside versus when it is sunny. But bringing an umbrella and cutting off my big toe do not. This isn’t actually an onerous restriction to satisfy if the states of the world do not impact the qualities of the outcomes—for example, in cases where events occur as a result of tickets being pulled from a hat. This method for revealing REU utilities suggests the following content for basic REU ‘ought’ judgments:

The REU judgment recommended by a graded desire for $\omega_i$: ‘To me (subjectively), for me (as grounds for my own action), what is the value contribution from efficacy at bringing about this outcome relative to efficacy at bringing about other outcomes holding fixed how the possible states of the world rank the available actions?’

An agent takes this content from desires, and his degrees of belief in the likelihood of outcomes, and then combines them with content from a third mental state governing how these limited value contribution judgments weigh against one another in contexts where the rankings of actions change. Because of the limited content of desire-based judgments on REU, these basic judgments do not form “contributing reasons” in the same sense as EUM. On REU the value contribution from a shot at a particular outcome might change given the lottery in which it is embedded. So with respect to the value contribution from individual outcomes at a given probability, the ‘contribution’ in question is not merely additive.

This brings into focus the consequentialist premise that Broome and Samuelson assumed and that REU denies. This particular brand of consequentialism is really the confluence of two points: that goodness is one-dimensional, and that the contributing value of efficacy at obtaining an outcome supervenes on the goodness of the outcome and nothing else.

With respect to the first point, on an internalist picture, we are forced to think of the goodness of a thing as the same as our strength of desire for it—or, in non-basic cases, as determined by the implications of our basic ‘ought’ judgments, which are themselves determined by strength of desire. One-dimensionality is just the idea that a strength of desire is simple enough to represent with just one number. It is one-dimensionality, whether applied to strength of desire or external measures of degree of goodness, that prevents degree of goodness or strength of desire from explaining how an agent could adopt a particular risk function. An easy way to see this is that a one-dimensional strength of desire gives you one number representing how good the outcome is. With a degree of belief, that gives you two numbers. But in order to figure out what to do an REU agent needs another independent number representing
the weight given to that outcome in this context. So far, an REU agent might agree with this—the assessment of the goodness of an outcome is one dimensional, and a risk function is determined through reference to other things. The conflict arises in the second claim.

The second point is the idea that the contributing value of efficacy at bringing about an outcome supervenes on the goodness of the outcome, or consequence, itself, and not on anything else. But because outcomes are just objects of desire, there is a bit of flexibility on this point. If an agent has a desire for insurance or an aversion to regret, for example, desires can make global properties part of a consequence.

REU seems to violate the second point—requiring that the value of efficacy at bringing about an outcome supervenes on both how good that outcome is and on global properties. One might argue about the details of whether we should call the confluence of these two points consequentialism or not. I would just like to note that the view could plausibly be called consequentialist, it patches the holes in the arguments from Samuelson and Broome, and it is, as I will continue to argue, correct.

REU also implies that ‘efficacy’ has a different definition holding fixed that instrumentally rational agents try to be as effective at satisfying their desires as possible. Both REU theorists and EUM theorists agree that efficacy at bringing about an outcome just is the probability at which your action will bring about that outcome. Where the disagreement about efficacy lies is in what this implies about efficacy at satisfying my desires in general. Take the following plausible Instrumentalist principle:

**Constraint on Instrumental Rationality (CIR):** A rational agent cannot choose an option that is, by his own lights, less effective than some available alternative at satisfying his desires.

Because REU and EUM agents disagree about what one ought to do, they will disagree about what the most effective path to satisfying one’s desires is. So insofar as they accept CIR, they are each pushed to give an account of **efficacy at satisfying desires** in accordance with their particular theory. Presumably, this definition will fit within a broader notion of efficacy generally. Some of the examples I will investigate in later sections help flesh out the implications of Independence violation for this concept.

---

5Buchak argues that REU is a consequentialist position as it is meant to present a theory where an agent only cares about consequences.
6.4 Independence and Group Decision

The next four sections analyze examples illustrating that the Independence axiom is essential to the judgments and practices that govern ideal reasoning.

A very expensive manned mission to Mars is taking place and you are mission commander. Your engineers have told you—and congress, the families of the astronauts, and the astronauts themselves agree—that backup rocket A is more effective at satisfying their desires than backup rocket B. Backup rocket A has a 99.99% chance of saving a mission and reaching Mars and a .01% chance of blowing up on activation. Rocket B has an 80% chance of reaching Mars and a 20% chance of leaving the crew stranded in orbit waiting for a rescue. Everyone agrees that in the event of a primary rocket failure, we would rather be operating with backup rocket A, and everyone has made this decision aware of the fact that there is no other relevant causal impact from pairing a backup rocket with a primary rocket. Now you have the choice, as mission commander, to pick the rockets you will use. Any primary rocket can be paired with any backup rocket.

Contrary to REU, no matter what the probabilities associated with the success of the primary rocket, it is rationally and morally unacceptable to pair it with backup rocket B. Indeed, if you ran the mission with backup rocket B on the basis that congress, the astronauts, and the families never weighed in on the pairing of the backup rocket and a particular primary rocket, then you would rightfully be “asked to resign”.

Notice here that one response from the REU theorist that isn’t advisable is to say that our reaction to this case is moral, and not rational, and that moral and rational demands sometimes conflict. The problem with an approach like this is that it risks the proliferation of conflicting sources of normativity. Neither EUM nor REU can be a complete theory of decision if there are sources of normativity that the theory conflicts with or simply doesn’t capture. On a view like this the recommendations of REU or EUM would come with the caveat, “unless there are other non-utility based considerations”, which isn’t the aim of either theory. Furthermore, on the interpretation presented in Chapter 1, we avoid this issue by way of the formal notion of desire. So whatever morality is and however it works, it should be captured as a consideration within a formal desire and, therefore, within EUM or REU.

---

Check appendix for demonstration that the REU of backup rocket A can be greater than that of B but that the pairing of B and a primary rocket can be higher than that of A and a primary rocket.

As I mentioned in previous chapters, I think that this implies that insofar as you want to endorse EUM or REU as a complete theory of rationality, your best bet for a theory of ethics is some form of contractarianism, e.g., (Gauthier 1986).
Not only are these claims about rationality and morality plausible, but they are connected with another plausible claim, which is that if I think that the efficacy of backup rocket A (at satisfying my desires regarding the mission) is greater than backup rocket B, then given that there are no other causally relevant facts, primary rocket C with backup rocket A will be more effective than the same primary rocket paired with B.

There are two ways to see this. The first is that if you asked for more of congress's time to weigh in on the pairing of the backup rockets with primary rockets, they would rightfully say that they had already answered that question for you when they told you that they preferred A to B. The second is that if you went ahead and spent millions pairing rocket B with a primary rocket, at your congressional hearing the common argument would be that a judgment about the pairing of A with any particular primary rocket was implied by the claim that “if there was a failure, they would rather operate with backup rocket A” and “that there is no other causal impact from the pairing of a backup rocket with a primary rocket”.

Of course, all of this is just to say that EUM, and the principles and concepts that go along with it, might be ingrainined in the minds of most people. But this is exactly my point. The real world plausibility of the idea that the REU-warranted inferences here would put someone in serious trouble suggest that REU employs concepts that people, in general, do not employ when they are considering ideal action guiding principles. If we want to allow that people, including senators, at least aim for rationality in situations like these, and that rationality is a psychological condition, then the fact that it seems so implausible that the group would accept REU justifications here is evidence that REU doesn’t capture what it means to be rational.

One might object that there are different rules governing individual action and action incorporating multiple people's preferences. In such a case, it might be possible that the rules of individual rationality recommend different choices than the rules governing these group decisions.

It seems to me that this can only be true to an extent. The rules that govern a rational group must be reducible to the effects of multiple agents operating as individual rational systems. This is true for the same reason that we wouldn’t want morality and rationality to come apart. If there are different norms governing group decisions, and they are not simply the implications of individually rational EUM or REU agents acting in a group, then EUM or REU cannot be complete theories of rationality.8

8This is a slightly different question from the one addressed in (Arrow). The problem there is that there is no consistent group decision procedure that captures all the intuitive standards we
This is not to suggest that there are no interesting puzzles of group decision making. But even if there are puzzles, there may still be some straightforward cases—like, for example, this case, where everyone agrees. In a case like this, it seems plausible that the job of the mission commander is to construct the rocket on the basis of the preference that everyone endorsed. He would be fired not for being irrational himself, but for failing to conform to the preferences of everyone with a stake in the mission.

But if we look at the norms of individual rationality implied by REU, then the information we have simply isn’t enough to know whether everyone endorsed the same preference, or how the balance of different preferences weighs out. If they are REU agents, then it is entirely possible that everyone prefers backup rocket A to backup rocket B and also all prefer the primary rocket paired with backup rocket B. The worst complaint anyone could raise against a mission leader who chose rocket B was that he proceeded without getting enough information about people’s preferences. But this doesn’t cohere with the natural prediction that people would have considered themselves to have already answered that question. Additionally, you could always assume that it was for some reason impossible to get everyone’s opinion on the second question. In this case, it seems to me that you will always be blameworthy for picking backup rocket B, and it is always within your rights to pick backup rocket A.

Given the way that the REU agent thinks about the world, there is a risk that placing an additional normative constraint within REU theory to avoid this behavior would be ad hoc, and not an implication of the mindset that an REU agent is understood to have. Without this additional constraint, the REU mission commander would be left wondering about everyone’s answers to the third question: to what extent are people willing to accept a chance at something worse for a chance at something better? This is an essential part of what it is to be an REU agent. I am arguing that the group would have already taken themselves to have implied an answer to that question, suggesting that they are, in fact, not REU.

Now, it may be the case that in reality people are observed to behave differently when they express group decisions from when they act alone. As far as the descriptive claim that people actually act differently is concerned, that could be for any host of reasons. But at least some of these reasons don’t imply that those agents are not still
EUM agents. For example, on the EUM interpretation I have presented, there are two reasons that immediately come to mind and suggest that rational agents would appear to abide by Independence more often when expressing group decisions rather than their own.\footnote{I say “appear” because I will argue that these agents obey Independence just as often in either case. It is just that representing the will of a group rules out small frequent preference changes that impact individual decision making—preference changes that can be easily misdiagnosed as Independence violation.}

Firstly, my desires are vague, but I might sometimes communicate them in a way that is artificially more precise. For example, I might say that the apple is worth twice as much to me as the pear. But, really, I’d be perfectly comfortable paying 2.00001 times the cost of the pear. When someone acts on my behalf, they may not take that vagueness into account, and, just to be on the safe side, pay only what I have precisely stated I am willing to pay.

Secondly, whenever I act I am free to redetermine what I want. I may notice, just a few moments later, I am more inclined towards the pear than the apple. This frequent minor preference change that occurs even outside the boundary of vagueness means that I may appear less EUM over sets of actions that are only relevant to me, than over sets of actions where I am attempting to comply with the will of multiple people. I don’t have the same ability to constantly check in with their desires that I have with my own. Even if their desires are as vague and variable as mine, my only option is to proceed on the basis of what they last told me they wanted. So a collection of my individual actions over time may seem to obey Independence in fewer cases, but that could be EUM if those appearances of Independence violation were just cases of preference change.

\section{Independence and Inferences about Efficacy}

Suppose you want to determine whether antibiotic A or antibiotic B is a more effective treatment—which one has a recovery profile you prefer. You have two trials. In trial A, patients are given a 75\% chance of antibiotic A, and a placebo otherwise. In trial B, patients are given a 75\% chance of antibiotic B, and a placebo otherwise.

My claim is that if you think that trial A was a more effective treatment for patients than trial B, then, insofar as you are raising the question, you have to think that antibiotic A was a more effective treatment than Antibiotic B regardless of the probabilities associated with convalescence on either antibiotic. Another way to put this is that it just seems built into the idea of “an effective treatment” that if you think trial A is more effective than trial B, then you have to think that antibiotic A
is more effective than antibiotic B. I do not take this to be a conclusive point on the issue—it is certainly possible that people could find this implausible. Indeed, REU theory does not require this. You would not be warranted in making that inference about efficacy because you might think that embedding the antibiotics in a lottery with a placebo changes the global properties in a relevant way \(^{10}\). But I do think that this is a consideration that many people would find constrained by their natural concept of efficacy.

We can start to investigate this example a bit more by digging into the principles that ground it and looking for those principles in other applications. My point is not that clinical trials actually work this way. My claim is that they could and that there would be no epistemological problem with it. What I think is so plausible is that if you had an argument that demonstrated that trial A was more effective than trial B, then that fact, plus the details of the construction of the trials, would be sufficient to ground the inference that antibiotic A was more effective than antibiotic B. The converse would also hold if you already had reason to judge that antibiotic A was more effective than antibiotic B and you wanted to make an inference about which trial would be better.

We can look at the implications of a general denial of this principle in other cases. You could just be interested in the science of antibiotics and want to publish an article arguing about the relative efficacy of these two antibiotics. But the more we raise the stakes the more obvious my thesis becomes. Suppose that this isn’t a clinical trial and that not everyone gets an antibiotic because there simply aren’t enough to go around. So everyone in the population faces a lottery of getting an antibiotic because you bring it for them, or getting nothing. If everyone prefers the recovery profile of antibiotic A to the recovery profile of antibiotic B and you can only take one with you to help treat an epidemic, I strongly recommend that you take A. This wouldn’t be required if you thought that you were on a mission to save REU agents. An REU agent might rather face a lottery with B in it over a lottery with A in it even if they would choose A in a context where there was no risk they might not be treated. So because not everyone will get some, if the population is made up of REU agents, they might all prefer that you take B instead. But if you could make enough, they would prefer that you manufacture and transport antibiotic A.

This just seems very strange. The fact that we cannot manufacture enough of an antibiotic does not change what antibiotic we should manufacture (assuming I can produce the same amount of treatments of either A or B). I wouldn’t, for example, provide A in a community that all had access to antibiotics and B in a community

\(^{10}\)There is a demonstration of the REU position in the appendix.
that didn’t all have access just because they didn’t all have access. I do not think that people would generally accept the justification that you brought B simply because not everyone was going to get some.\footnote{Even though this is a case of saving lives, it isn’t necessarily a case of a group decision the way the Mars mission is. In this case, you could be expressing your own desire to save lives, not acting on the will of the victims. But, even in a case like this, other people might evaluate the appropriateness of your reasoning process.}

One might object that in a given situation the recovery profile of A might not make as much sense as the recovery profile of B. For example, maybe A results in more deaths than B even though almost everyone gets better. These deaths, in the context of the ongoing epidemic, might stretch emergency services too thin, causing even more problems. In such a case, we need to weigh these deaths differently than when considering the antibiotic outside of the context of emergency response.

My reply is that these other problems brought about by the number of deaths change the decision problem. Now, there is an additional utility consideration. What REU suggests is that the simple fact that the antibiotic is now embedded in a lottery can change whether I ought to pack antibiotic A or antibiotic B. But it seems to me that it would be a difficult argument to justify that reasoning to other people who wondered why you brought B even though everyone judges A to be a more effective treatment. If there were REU agents as a substantial portion of the population, our intuition should be that an argument like this wouldn’t be out of the question if you were asked on the news about your relief effort.

Something that this example begins to illustrate and which will come up in later examples, is that Independence violation and justification for it are not as general as REU seems to suggest. If one were inclined to manufacture a different antibiotic in a community where not everyone could get treatment, the consideration that makes this the case would not be the simple fact that not everyone could get treatment. But this seems to be what REU suggests. For an REU agent, the same considerations apply to this case as apply to the Allais Paradox. In either case, it is the risk function that determines the weight of the reason. But intuitively, it seems that the justification for manufacturing one antibiotic or the other in this case must be of an entirely different kind than the decision to choose \(l_3\) over \(l_4\).

Again, the point of this example is to demonstrate how unnatural judgments using the REU notion of efficacy are, not that they are impossible. These examples are meant to suggest that REU and the collection of concepts implied by it just aren’t what people are talking about when they are talking about ideal practical reasoning.
6.6 Independence and Sequential Choice

There are at least two different ways that violating Independence can create problems of long-term efficacy for an agent. One has to do with the way that an agent values individual smaller goals in light of larger long-term goals. The other has to do with how risk sensitivity impacts how an agent chooses when he faces a lottery all at once versus when he faces its component parts at different times. In both cases, the fact that the formal notion of desire forces an agent to choose what they prefer at that time puts agents who deviate from EUM on a course that distances them from their long-term goals. This happens, perhaps unsurprisingly, because weighting outcomes differently from their expected utility leads to actions that end up maximizing something other than utility over the long run.

6.6.1 Long Term Decision Making

Long run decision making raises difficult question for both EUM and REU theory. In a sense, the considerations of this section do not obviously tip the scales in favor of either REU or EUM. However, there are two reasons to investigate the subject a bit here. The first is that the psychological interpretation I’ve provided yields a Humean account of long-term decision making that contrasts with what one might have thought about sequential rationality. Looking at sequential decision making on REU and EUM given my interpretation can help highlight the costs and benefits of this approach.

Long run decision making has traditionally been an interesting subject because an EUM agent might change what he would effectively pursue when the outcome is gotten all at once, versus when it is acquired piece-by-piece in smaller sequential decisions. As I will discuss in more detail momentarily, this inconsistency arises in the EUM agent because how he values more of a good changes the more he has. An REU agent can have the same kind of long term inconsistency for the same reasons. But being REU opens up another opportunity for sequential inconsistency that isn’t present for the EUM agent. Specifically, the REU agent’s sensitivity to risk can cause him to change what he effectively pursues if an outcome is gotten all at once versus piece by piece. The way this plays out is an implication of REU that might not have been immediately obvious from the formal definition or the psychological interpretation. So, in this sense, questions of long run efficacy bring to bear new considerations that might weigh on a person’s acceptance of either theory.

Here, I’ll raise an example that is similar to those in Buchak chapters 6 and 7. However, this example is different in two respects. The first is that it explicitly, rather
than implicitly and without argument, restricts the agent in question from having
the kinds of attitudes that allow an EUM agent to have the aforementioned sequen-
tial inconsistency, leaving open only the REU inconsistency. Secondly, it operates
against a backdrop of the psychological interpretation I’ve presented—a psychologi-
cal interpretation that restricts the kinds of responses theorists might make to cases
like this.

There are two different rifles, rifle A and rifle B. As a competitive marksman, you
need to choose a rifle each time you participate in a riflery competition over your
career. All you care about is how many medals you will win. Introspecting on your
desires, and what preferences you have over simple lotteries or straightforward trades,
you determine that \( u(\text{Gold Medal}) = 4, u(\text{Silver Medal}) = 2, u(\text{Bronze Medal}) = 1 \).
You also conclude that you value medals linearly, by which I mean that \( u(2 \text{ medals}) = u(\text{first medal}) + u(\text{second medal}) \). Being an expert in guns, and your own skillset,
you determine that rifle A gives you a 60% chance of winning gold and a 40% chance
of bronze. Rifle B guarantees that you will win silver every time. Furthermore, this
expectation has nothing to do with your choice of rifle impacting the matches or
any expected change in the competition over time—that is, all trials are indepen-
dent of one another. Over the course of your career, you expect to compete in 100
competitions. The question is “what rifle is it rational to choose?”

A natural thought in this case is that if a rifle is more effective over the long
run, then it will be more effective in each individual trial. The converse should also
be true—I can’t be more effective each individual trial, and then end up being less
effective in the long term.

What should happen then, is that given the problem specifications, the agent
should choose rifle A every time. Over the course of his career, rifle A will lead to
approximately 60 gold medals and 40 bronze medals, with a utility of 280. Rifle B,
on the other hand, will lead to 100 silver medals with a utility of 200. So, at the end
of a career, the marksman will wish he had chosen rifle A\(^{12}\). It seems, then, that a
theory of rationality should recommend choosing A in each individual instance.

While EUM does recommend choosing A in each individual instance \((.6(4) + .4(1) > 2)\), REU doesn’t necessarily require that. As long as \( r(.6) < 1/3 \), choosing
Rifle B will seem preferable in each individual instance even though the mix of medals
has a higher REU than 100 silver medals\(^{13}\). Given the formal notion of desire, the
REU agent will choose rifle B each time and then get to the end of his career wishing

\(^{12}\)If such a thing were possible, it seems consistent with both EUM and REU that at the end of
the contract he would pay to switch his past choice from B to A, or the payoff profile from long
term B to long term A. This kind of consideration will arise in the next example.

\(^{13}\)The REU calculation is in the appendix.
he chose rifle A, or at least had the results of rifle A. This happens because he weights medals differently at the time of choosing—as a possible outcome of his action—than how he weights them once he has them, or as a component of a larger collection gotten all at once.

But being effective in the short term and being effective in the long term don’t always go hand in hand on EUM either. An EUM agent might have decreasing marginal utility in money, for example. As a result, he might be risk-averse in each individual decision. This could lead to choosing an option in each case that has lower expected value long term than a riskier alternative. So if there is a problem specific to the REU agent here, we need to be clearer about it.

To get clearer about the importance of this example, we can look at short term and long term utilities against the backdrop of the psychological theory I present. Usually, when we construct extensive form decision problems, the only outcome with a utility assignment is the terminal outcome. In this case, that would be a collection of medals at the very end of your career (or the total amount of money the risk averse EUM agent would end up with). On a picture like this, there is no sense to be made of the idea that I am effective at satisfying my desires in an individual competition because there is no outcome associated with one competition. The outcome is what happens after 100 competitions.

But this approach is unnatural and unnecessary. If we expand this principle out to an agent’s entire life we can see why. If outcomes are only what happen at the very end of a decision tree, then why not model an entire human life as one decision tree? Going to college isn’t a terminal outcome. It is a choice on the path to the terminal outcome, which is whatever life I end up living. If we were forced to model agents such that the outcome was always at the very end of the tree, it isn’t clear why capturing the agent’s entire life wouldn’t be the strictly correct approach. But this approach is inconsistent both with common practice and with interpreting formal outcomes in terms of objects of desire. With respect to practice, often, we just think about the decision at hand and treat it as having its own outcomes. Every time I act I do not think of the outcome at hand as part of the overall story of my life. But then again, maybe I should. The problem with this claim is that it is not clear why I should. If my objects of desire are just simple things like, ‘having an apple’, or ‘going to college’, then according to the TF story I am thinking of those as ends, not means to some bigger overall picture. If I am free to do this, and if outcomes represent my objects of desire, then I am free to be the kind of being for whom it would be appropriate to model me as having intermediate outcomes in my decision tree on the way to the terminal nodes.

One might object to intermediate outcomes by arguing that it assumes that agents
reason *myopically*, failing to look at the long term results of their short term aims. But I’m not assuming here that all agents reason myopically, I’m just noting that the interpretation we have on hand for either REU or EUM doesn’t rule it out as irrational. An agent could, if they wanted to, frame the objects of their desire in terms of short term proximate goals. This is in line with the quasi-Humean internalist approach I’ve been presenting. What this means is that sophisticated choice, understood in terms of backward induction itself, or reasoning from the expected outcome at the terminal node to the expected utility at previous branches, is a constraint that only applies after an agent sets his outcomes. If an agent thinks in terms of intermediate outcomes, then there is nothing in the model interpretation that suggests sophisticated choice would apply. The agent just does at a time what he has an all-things-considered formal desire to do at that time—in other words, he aims at the object of desire, or outcome. If he thinks of the object of desire in this short term sense, then that is all there is to the story. Events that happen later on do not impact his choice at this node unless they have been incorporated into his considered objects of desire—a condition we are assuming doesn’t obtain in this instance\(^\text{14}\).

Buchak and anti-Humean EUM theory claim the opposite, “When deciding what to do now, an agent needs to take into account her expectations about the future. If she expects to take two copies of Z in the future, she will prefer Z now, and if she does not, she will prefer W now” (Buchak 231). What I’m claiming is that neither the interpretation of REU nor the interpretation of EUM explains the presence of this ‘need’. On either theory, you are free to reason myopically. If you reason myopically, you are perfectly free to pick W now regardless of what you’ll want later.

With respect to this individual example, it seems that happiness with the long run results of rifle choice *could be* the result of the aggregate happiness with each individual competition, or it could be something the agent thinks of differently. If the agent took an approach of aggregating the value of individual competitions, then what he wants to end up with in the end will always be compatible with what he

\(^{14}\)One might think that this is an implausible story in virtue of the fact that it doesn’t capture prudence—that I shouldn’t just pursue what I want now, but that I should take care now to facilitate the satisfaction of my future desires. The problem of prudence is a problem of the Humean picture itself, and not appropriately handled through the purely formal constraint of putting utilities at the end of decision trees. If we want to explain the rational appeal of prudence, we need to do that within the substantive psychological theory, and do so in a way that doesn’t fall victim to a ‘So What?’ objection that would plausibly arise asking an agent to do something to satisfy desires he doesn’t have yet. I think that the rational appeal of prudence is best handled by way of making desires a rational response to something else—as I suggested might be the case in chapter 2. This way, my pleasure later could be just as much a reason for a present desire as my pleasure now. This could explain prudence while avoiding a ‘So What?’ objection. None of this changes anything about the example at hand.
wants in each individual instance. But if he doesn’t aggregate in this way, his short term focus could lead him to a situation where what he wants in each individual instance leads to a terminal outcome that is different from the one he would prefer if he chose it all at once. In this particular example, that could amount to something like preferring a gold medal in each instance and yet also preferring a career that started off rocky and rallied to successive dominant performances.

Buchak points out that an EUM agent might prefer something different in each individual trial of a repeated decision problem like this, and yet prefer an incompatible terminal outcome. For example, if an EUM agent has decreasing marginal utility in money, then in each individual trial he will prefer a decision that takes less risk than the decision that would maximize his expected monetary payoff over the long run (225). What this shows is that the incompatibility of long and short term aims is not enough in itself to condemn REU over EUM.

But this isn’t an example about the long term incompatibility of long and short term goals in itself. Rather, this is an example of long term and short term incompatibility in an agent that supposedly values the goods in question linearly—which, on REU or EUM, he should also be free to do. The reason an EUM agent’s actions might not maximize expected monetary value is if he doesn’t value money linearly \(u(\text{the money I have})+u(\text{the money I'll get}) \neq u(\text{the sum of those amounts of money})\). An additional dollar could be worth less to him the more money he has, breaking linearity. In a case like this it wouldn’t be surprising if he had an incompatibility between what he would choose in each individual instance and what he would choose if he picked the terminal node all at once because his attitude about the value of additional money changes as he gets more. So as he goes through the individual decisions that make more money, he is willing to accept less and less risk as that process goes on.

What one might think is strange about the REU alternative here is that the agent’s attitude towards the relative value of money and risk hasn’t changed over time. He values medals linearly and yet his decisions have taken him along a trajectory away from the long term best outcome. What causes the REU agent to bend his behavior away from his long term goals is how he views the risk associated with one match versus the risk associated with repeated action. The more competitions, the smaller the risk Rifle A leaves you worse off than Rifle B. The REU agent is sensitive to this, and changes his choice depending on whether he faces the equivalent of many trials, or just one trial assessed myopically. So there is an additional opportunity for sequential inconsistency that the REU agent faces that the EUM agent does not. Specifically, he will have an incompatibility with his short and long term goals even though he values the outcomes the same way over the long-term.
Thinking that the REU agent’s position here is strange seems to me to just be another expression of the consequentialist supervenience principle I mentioned earlier. Valuing intermediate outcomes linearly in this sense is to have a stable attitude about how good it would be to get more of that item—in this case, more medals. One might think that if my attitude about this remains stable, then the contributing value of each individual trial would remain stable regardless of what other trials might happen. For some, when an EUM agent fails to maximize expected monetary value this is less surprising because his attitude about how good it would be to get an additional dollar changed over time.

Of course, none of this is to suggest that the REU approach is impossible. I just want to note that this is something someone could legitimately find strange about REU and yet there is no corresponding problem in EUM.

Just to clarify, I am not claiming that an agent must think in terms of intermediate outcomes, or that he must treat the value of the long term outcome as the aggregate of those intermediate outcomes. Rather, I am claiming that on either REU or EUM, the psychological picture I have presented implies that an agent is free to think in terms of intermediate outcomes and to value the long term outcome as the aggregate of the short term outcomes. This is the sense in which this argument is different from those in the literature, and the first half of what ultimately makes it work. Buchak points out that we can’t just assume that agents value long term goals by aggregating short term goals:

I will show that [any argument about hypothetical long term choices] begs the question against the globally sensitive agent. In particular, the argument assumes that if an agent is willing to pay $x$ for one gamble and $y$ for another then she is willing to pay $x + y$ for both, and this assumption is generally false for the globally sensitive agent. Furthermore, we can’t assume there must be a context in which bets are cumulative in this way unless we start with the assumption that global sensitivity is irrational (216).

I’m not sure that I agree that such an argument would beg the question, but I do agree that it would be flawed. Specifically, it would assume something that is unjustified, which is that the value of bets is cumulative in this way. The difference in this argument is that we are not assuming that this kind of aggregation must be true for rational agents. Rather, I’m arguing that neither REU nor EUM rules it out, and then asking what would happen for agents like this who are EUM or REU.

Of course, all of this assumes resolute choice is not an option. This is the second half of what makes this example work, and how it differs from arguments presented
elsewhere. If the agent could choose rifle A as a plan and stick to it, then he would work his way towards the preferable collection of medals. But, as I argued in chapter 1, resolute choice isn’t an option at all unless it is understood as a case of preference change. In this particular instance, the REU agent’s perspective includes no pressure to change his preferences. Before the first competition he recognizes that he would prefer rifle B in each individual competition, and he simultaneously recognizes that he would prefer something different from the long term expected outcome of B in every competition. He also claims that the value of the outcomes of the trials are independent of one another—so more medals do not become less valuable to him the more he has. One might think this is strange because he acted in accordance with his unchanging attitude about the value of more medals to a long term outcome that he expected to disprefer to an available alternative. But nothing about REU theory itself suggests anything about his reasons that imply something wrong with this approach at any point along this path. So he keeps the preferences he has, and resolute choice doesn’t come up.

Despite the term ‘myopic’, the interpretation of neither theory provides a reason for thinking that there is something wrong with conceptualizing outcomes in terms of short term desires. So questions of long term sophisticated choice and backwards induction do not necessarily apply. Additionally, in cases like this one, neither theory explains why myopic agents would see anything wrong with the incompatibility of their short and long term goals. If an EUM agent has decreasing marginal utility in medals, his long term and short term decisions will be inconsistent as well. The point is that there is an additional way for the REU agent to wind up in sequential inconsistency, and resolute choice is not a solution for anyone.

### 6.6.2 Valuing Lotteries by their Parts

This sequential choice example comes directly from *Risk and Rationality*, but we now have some additional tools for investigating its implications. The puzzle arises when considering extensive form variants of the Allais problem. For convenience, here is the original problem again:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2-11</th>
<th>12-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$0</td>
<td>$5M</td>
<td>$0</td>
</tr>
<tr>
<td>2</td>
<td>$1M</td>
<td>$1M</td>
<td>$0</td>
</tr>
<tr>
<td>3</td>
<td>$0</td>
<td>$5M</td>
<td>$1M</td>
</tr>
<tr>
<td>4</td>
<td>$1M</td>
<td>$1M</td>
<td>$1M</td>
</tr>
</tbody>
</table>

173
We can think of the choice between $l_3$ and $l_4$ before the agent knows which ticket has been drawn, and after the agent knows which ticket has been drawn.

![Figure 6.1: Allais Decision First](image1)

![Figure 6.2: Allais Chance First](image2)

We can take the case of an REU agent who preferred $l_4$ to $l_3$. It is entirely possible that this agent would prefer $l_4$ to $l_3$ without knowing which ticket has been pulled, and then, after learning whether the ticket is between 1-11 or 12-100, strictly or weakly prefer each part of $l_3$ to each part of $l_4$.

This isn’t actually strange in a new way—it is all well within standard REU behavior. Even though he prefers all of the parts of $l_3$ over the parts of $l_4$, putting those parts together changes the overall risk profile. Any strange implications in his efficacy concept at this point would just be variants of what I have already discussed.

Additional trouble arises when we put the two decisions together. Assume that the preference for $l_4$ over $l_3$ is strict enough that the agent would pay $1 to face $l_4$ rather than $l_3$. Given sophisticated choice, if the REU agent faces the right-hand decision problem, his practical stance creates a requirement to choose something extensionally equivalent to $l_3$. But if he faces the left-hand decision problem, he will choose $l_4$. So given his strict preference for $l_4$ over $l_3$, he should be willing to pay $1 to face the left-hand decision problem over the right-hand problem. This can be represented in Figure 6.3.

What is strange about this is that if he had just chosen the parts of $l_4$ from the right-hand problem, he wouldn’t have needed to pay the dollar, and he would have been able to assure himself payoffs that were better (by $1) than his alternative strategy regardless of what ticket was pulled.

---

15Given these preferences, REU requires the following: $u(\$0) + r(10/11)(u(\$5 \text{ Million}) - u(\$0)) \geq u(\$1 \text{ Million}) \geq u(\$0) + r(99/100)(u(\$1 \text{ Million}) - u(\$0)) + r(10/100)(u(\$5 \text{ Million}) - u(\$1 \text{ Million})$. One way this holds is if $r(10/100)$ is close to zero.

174
Figure 6.3: Paying to Avoid Information
The problem with this particular accusation of strangeness is that, consistent with the reasons for sophisticated choice, the parts of \( l_4 \) in the right-hand problem simply aren’t available. He knows what he will choose when he gets to time \( t_2 \), and the parts of \( l_4 \) are not it. It is not as if he is paying $1 for \( l_4 \) when he could have had \( l_4 \) for free, \( l_4 \) was never really an available option. Our question, then, is about whether the preference for \( l_4 \) overall is compatible with the preference for the parts of \( l_3 \) (Buchak 199-200).

But this example does highlight the way in which it is a bit counterintuitive to call the REU agent a consequentialist. If all this agent was interested in was outcomes and how effective he could be at making those outcomes obtain, then he would be free to make the right hand problem just as attractive as the left hand problem by giving himself a sure shot at $1 million either way. That is, preferences aside, he is capable of choosing the equivalent of \( l_4 \) in the right hand problem. So if his preferences were determined solely by the quality of the outcome and the probability at which he could make it obtain, then he would be indifferent between the right hand problem and the left hand problem. What this highlights is that the REU agent isn’t interested so much in outcomes themselves, but in risk profiles—facing the prospect of an outcome or set of outcomes at a particular probability. That isn’t in itself an objection to REU, but it will be useful later on in presenting an error theory of reflective Independence violation.

Buchak argues that there is nothing irrational about preferring the parts of \( l_3 \) to the parts of \( l_4 \) while preferring \( l_4 \) to \( l_3 \) overall. The argument begins by noting that whether you are an EUM or REU agent, certain pieces of information might be misleading—they can be true and at the same time lead you away from doing what would actually be best. Buchak gives the example of hearing on the radio that “there is traffic on the bridge” but not knowing which bridge. This information could rationally impel a decision to take the bus, rather than drive, by increasing your degree of belief that there is traffic on the bridge you actually want to take. So if the traffic turns out to be on the other bridge, then rationally responding to the information you received led to a lower utility outcome than was actually available (203).

But whether you are REU or EUM, whether information has positive value or not will depend upon the perspective of evaluation. In the Allais case, the misleading information is that “a ticket from 1 to 11 was drawn” in the event that ticket 1 was drawn. This would motivate the REU agent to choose a part of \( l_3 \) even though it would turn out worse than the alternative. What is different about the REU and EUM agent in this respect is that, for the EUM agent, if the risk that information might be misleading is enough to motivate a choice of a part of \( l_3 \) after getting
the information, it is enough to motivate a choice of $l_3$ over $l_4$ from the beginning. Because this is not the case for the REU agent, there can be a kind of *instrumental disvalue* to information where, from perspective of $t_1$, it makes sense to pay to avoid ever getting that kind of information. In the case of EUM, even if information is occasionally misleading, it will always be at least weakly preferred to have it. In the bus case, for example, getting information about there being traffic on one bridge or the other would, in the long run, turn out better for you even if it was misleading in some cases\(^{16}\). This means that the expected utility of the information is higher, so EUM would suggest taking it.

Buchak takes this to be the primary consideration against adopting REU as a theory of rationality. The idea that a superset of the true information presented to an agent could be dispreferred seems strange, but there is nothing wrong with it from the perspective of the REU agent (210).

I agree with Buchak’s analysis of this situation. I don’t see any reason why, from the REU agent’s perspective, all information has to be valuable, or that a superset of information must be at least as valuable as a subset. In fact, this seems to naturally match the idea that there is no normative view from nowhere—that everything is good or bad only from within a teleological functionalist perspective. So the fact that a superset of information is not always weakly preferred is not even a reason, by itself, to reject REU. Rather, the reason this consideration weighs against REU is just that it doesn’t accord with our normal intuitions about good practical reasoning, and as such is a reason to think that we aren’t like that.

We can consider the strangeness of the instrumental disvalue of information as a strangeness already accounted for in the debate between EUM and REU. I want to raise a different concern, which is that the REU agent takes $l_4$ to be more effective at satisfying his desires than $l_3$, while at the same time thinking that the parts of $l_3$ are more effective at satisfying his desires than the corresponding parts of $l_4$.

Here is a natural thing to think: if $l_3$ is more effective at satisfying my desires when tickets 1-11 are drawn, and when tickets 12-100 are drawn, then $l_3$ is more effective at satisfying my desires overall. REU doesn’t capture this, and is made more alien as a consequence. This is the conceptual analog in efficacy of NIE from chapter 3. An example similar to the previous cases could be constructed to dramatize this point, or to present a valid-seeming inference from the efficacy of the parts to the efficacy

---

\(^{16}\)Resolute choice understood in terms of preference change will not resolve this issue. This is because, from the perspective of the REU agent, there is nothing wrong with the way he expects himself to view things at $t_2$. His being REU includes the idea that he endorses, at $t_1$, the way he will think at $t_2$. There is no normative pressure from within his perspective to change anything about his preferences or the way he is approaching the problem.
of the whole that isn’t warranted on REU, but I will skip this step.

6.7 Independence in the Real World

If everyone is an REU agent, or even if a substantial portion of the population are REU agents, we should be able to find evidence of their existence in the world. This evidence wouldn’t just be in the types of decisions people tend to make, as that could be the effect of predictable forms of irrationality. Rather, the best place to look would be in the kinds of practices we engage in that are supposed to facilitate ideal decision making. If there are real world practices or standards that assume people obey Independence, for example, then we would expect REU maximizers to object to those practices on the basis that they distort outcomes by failing to naturally map onto considerations governing the comparison of reasons. If no one raises objections like this, and the practice is sufficiently important and widespread, then we have reason to think that people aren’t actually REU agents.

As it turns out there is at least one popular and important practice that assumes that people obey Independence—primary elections. Members of political parties must vote for the single candidate that they want to represent their party in the general election. Assuming a two party system, this means that an agent must vote for a candidate from his own party that he wants to participate in a general election against a candidate from the opposing party. That candidate would also have been chosen by primary election, but one individual is only allowed to participate in the primary election for one party. It seems reasonable to represent a particular president as a lottery. A given president might return any number of outcomes capturing things I actually care about—general welfare, national security, etc. So my choice of candidate A is not a choice of a single outcome, but of a particular lottery—50% chance of increased national defense, 30% chance of improved social programs, etc. And a general election, from my own point of view, is definitely representable as a lottery (there is some chance the candidate from my party will win, and some chance the candidate from the other party will win). So a primary election is a situation where an agent is forced to make a choice about what sub-lottery (candidate) they want as part of a compound lottery (a general election) before she knows what the other lottery composing the compound lottery (the opposing candidate) will be. In formal structure, this is nearly identical to the Mission to Mars case, where agents needed to choose a backup rocket before choosing a primary rocket. But it is also clearly representable in terms of the Independence Axiom, where \( l_1 \) and \( l_2 \) are primary election candidates from my own party and \( l_3 \) is whatever candidate from the other party is selected.
As an REU maximizer facing a primary election, I have a problem. The candidate from my own party that I want win might be different from the candidate from my own party that I want to be part of the general election\textsuperscript{17}. I might think that the reasons in favor or against a candidate weigh against one another differently when she is running against a particular opponent, and that given this change in the weighting, I don’t want to vote for her in the primary election.

Now, wanting to vote for a different candidate in the primary election from the one you would want to win the presidency is not in itself odd. People do it all the time. The key consideration here is why they do it. There are two plausible reasons one might vote for a different candidate in the primary election from the one you would want to win the general election. Only one of these reasons is a frequent public complaint about primary elections and both are compatible with EUM. The natural REU complaint against primary elections just doesn’t come up.

The complaint that people actually bring up against primary elections, and one of two reasons why one might vote for a different candidate in a primary than the one you would vote for in a general election, is the problem of determining electability. This can happen because some candidates’ strengths serve them better in opposition to specific competitors. So, for example, I might think the military officer has a better chance if he is running against the businessman than if he is running against the career politician. Another reason I might change my vote on the basis of electability is if I think the general population is much more moderate than I am. If the candidate from my own party that I would want to win the general election is too liberal to plausibly win the general election, in the primary I might vote for a more moderate candidate as the lesser of two evils.

The problem with this explanation as a motivation for REU is that it doesn’t violate Independence. Electability is the probability at which a particular candidate will be elected in the general election. So the choice between $l_1$ and $l_2$ changes the probability of $l_3$ in this case. This is both the complaint that people usually raise against primary elections, and perfectly compatible with EUM.

The second problem people might have an issue with primary elections they don’t bring up publicly as much, and perhaps for good reason. This is the case of there being emergent reasons for preferring one candidate in the general election over another. For example, suppose I have the opportunity in my primary election to vote for a white woman or a white man. And the opposition candidate will either be a black man or a white man. One might think that having a female candidate win over a white male candidate reflects something positive about social progress that is

\textsuperscript{17}Like the Mission to Mars case, the backup rocket I want to be operating with if it comes to that might be different from the backup rocket I want paired with a primary rocket.
not reflected in having a white female candidate win over a black male candidate. So for that reason, the candidate that I would want to be president might be different from the one I would want to win the primary election.

The problem with this as an explanation for dissatisfaction with primary elections is that it is both mildly troubling and perfectly compatible with Independence. People don’t bring up considerations like this in elections not just because of stigma against overt identity politics, but because it not the right kind of reason to vote for someone. Voting for someone in good faith is meant to be an expression of your belief that this candidate will produce the best outcomes through their governance. If you vote for someone because it would be good if they win, and not because they would be a more effective president, then that seems a bit distasteful. So even if people actually have considerations like this, there are reasons why this wouldn’t be the most common objection to primary elections in the public discourse.

But more importantly for our purposes, emergent reasons are perfectly compatible with EUM. Here we have a situation with a need to redifferentiate outcomes. If an election represents a triumph for social justice, and that is a reason to prefer a candidate, then each outcome that occurs as a result of that lottery gets a premium for including the property ‘and resulted from an election which was a triumph for social justice’. So this complaint against primary elections, while plausible, is not a complaint issuing from a violation of Independence.

REU agents, if they are out there, would have a very particular kind of complaint about primary elections. They would say that without there being any emergent reasons, and without there being any difference in electability, the design of primary elections causes a change in the way that reasons for a candidates weigh against one another. So while the relative value of things like general welfare and national security did not change, and the efficacy of each candidate at bringing those about did not change, putting two candidates in a general election together changes the way that a given chance at general welfare weighs against a given chance at national security. You might find this remarkably strange, or you might not. But what we should all agree on is that if REU agents are a substantial part of the population it is strange that this complaint isn’t part of the public discourse. When I turn on the news during election season I don’t see complaints that translate into objections to the Independence axiom. This is a reason to think that REU agents, if they are out there, are not numerous.
6.8 On the Tendency to Violate Independence

One of the takeaways of the primary elections example (and perhaps the antibiotics example) is that Independence violation is context dependent. Agents do not have general attitudes towards risk that aim at uniform sensitivity to global properties across all decisions. This is also reflected in experimental results. There are decisions where people seem to obey alternatives to EUM—like the Allais Paradox, or other cases motivating Prospect Theory. And there are cases where people seem to obey EUM. This situation is such that none of the non-traditional alternatives to EUM capture human behavior more accurately in a sufficient number of cases to have a clear lead in the game of representing human action (Ross 2005, 175).

There are various causes for mistakes regarding what one ought to do in a particular situation. Sometimes, context can make a difference. The Lenny case from chapter 2 is a good example. Lenny misjudged his evidence and came to a false conclusion about what his desires actually were. He did this in the context of his overwhelming trust for George. People also might miscalculate the probabilities associated with outcomes, or misjudge the degree to which a career in the military would lead to happiness, etc., depending upon the context in which the question is asked. Because beliefs play such an essential role in the Teleological Functionalist picture of practical reasoning, all the standard mistakes people make with doxastic reasoning can occur in the context of practical reasoning as well.

But noncompliance with EUM is not just about individual decisions. It happens at the level of theory choice as well. People who know what they are talking about, Buchak, McClennen, etc., reflectively reject EUM. That is, they think acting contrary to Independence is what they ought to do or what their reasons recommend.

There are two reasons to think that the doxastic errors that motivate the above mistakes in individual decisions don’t explain the choices of professional decision theorists who reject Independence. Firstly, professional theorists have had enough time reflecting on questions to avoid doxastic errors of these kinds. It isn’t plausible that over the course of an entire career, a professional decision theorist is just making a mistake multiplying probabilities, for example. Secondly, the content of a claim about REU theory is different from the content of a claim about a particular action. Buchak isn’t just making a claim about what she ought to do when faced with the Allais Paradox, she is making a claim about the nature of reasoning itself. So, if REU theory is incorrect, then her error doesn’t regard the exercise of rationality so much as determining what the standards of rationality are in the first place.

I will call the problem of explaining why agents might think that reasoning recommends Independence violation the problem of *reflective* irrationality, or reflective
Independence violation. So this is the problem of explaining not just how people could make a mistake in the exercise of reasoning, but how people could make a persistent mistake about the nature of reasoning itself. Because informed parties disagree about the nature of reasoning, any true theory of rationality needs to be compatible with the ultimate explanation of that disagreement. Because people predictably violate Independence, it is plausible to think that theorists who reject Independence are tracking a real influence in human decision making. Explaining reflective irrationality might tell us something about the relationship between rationality and whatever deliberative tendency these theorists have glommed onto.

As it turns out, Teleological Functionalism places some strong restrictions on the possible explanations of reflective irrationality. In this section I want to investigate these restrictions, and whether a plausible explanation for reflective irrationality can exist within them.

Figuring out why people reflectively endorse Independence violation is a remarkably difficult question, one that deserves its own independent investigation that I do not have the space to sufficiently explore here. There are at least two reasons why answering this question is so complex. Firstly, notice that all plausible candidate theories of rationality, those that endorse Independence and those that don’t, must make room for the fact that people are predictably irrational. People predictably mess up the Wason Card Selection Test, are susceptible to framing effects, base rate neglect, etc. Neither EUM nor REU warrant this kind of behavior. So whatever psychological considerations govern rationality, there must be some psychological/neurobiological story to tell about what causes all these mistakes. Giving an accurate characterization of the network of mental states and/or influences that leads to predictable irrationality is potentially just as great a task as presenting a characterization of the rational states. An inquiry like this should plausibly rely upon evolutionary, experimental, and introspective evidence that goes beyond the scope of this work. Supporting a theory of irrational influences would be the best way to provide an error theory for reflective Independence violators. If the same types of influences govern framing effects and REU decisions, for example, REU theorists might be misdiagnosing irrational motivators as rational ones.

But providing an error theory for reflective irrationality faces a challenge that is specific to Teleological Functionalism. The objection is that, on TF, any system that resembles rationality enough to create confusion must itself be rational. In the following subsections I will go through this problem in more detail and several possible solutions to it (3 that fail and two that might work). While this will not provide a completed error theory for reflective Independence violation, it is a starting point that would ultimately be judged against the results of an extended investigation.
of irrational influences. Ultimately, TF would need to be compatible with the results of that project.

6.8.1 Objection: TF and the Problem of Reflective Irrationality

Teleological Functionalism seems to rule out the most straightforward approach to giving an error theory of Independence violation. That is, it would be easy to think that people who reflectively endorse Independence violation confuse the demands of another decision making system with the demands of rationality. For example, the neurobiological characteristics that motivate Hulse, Read, and Schroeder, or the behavioral evidence that motivates prospect theory, seem to suggest that if rationality exists at all then it must exist alongside another system that has a regular and predictable impact on what people choose to do. One way to explain the reflective endorsement of EUM violation could be to suggest that people are confusing the demands of this system with the demands of rationality.

But it is not clear that Teleological Functionalism still works in the context of this kind of story. If people are confusing the demands of the irrational system with the rational, then TF seems to suggest this is because the demands of these systems present themselves similarly from the first-person point of view. As I’ve described TF, it is about having mental states with first-personal aim properties. But whatever states govern prospect theory, those seem to be about first-personal aim properties too. For example, they incline me to form a judgment that I ought to aim at avoiding loss to a greater extent than seeking gains. But if all that deontic normativity amounts to on TF is having a mental state that has these kinds of aims, what makes the mental states I described in chapter 2 deontic normative and those that govern purportedly irrational behavior not deontic normative?

6.8.2 Failed Reply 1: Coherence with Other Concepts

One approach to answering this question is not open to instrumentalists. That is, you could say that the ‘rational’ system coheres with our other normative concepts—concepts about efficacy and the way the value of parts of lotteries or collections of actions contribute to the value of the whole. But this would make rationality about coherence as such. And for that reason, it would be subject to a ‘So What?’ objection—“I know I have mental states that aim at EUM, but I also have mental states that aim at prospect theory. So why should I care that the EUM states cohere with other concepts I have?”
6.8.3 Failed Reply 2: The Existence of Actual REU Agents

The approach I’ve taken in this project has never been to suggest that REU agency is impossible. I’ve just argued that REU agents are strange or uncommon. So even if REU maximizers don’t make up a large portion of the population, and even if they seem strange to EUM agents, one possible explanation for a reflective endorsement of an Independence violation is that the person in question is actually an REU maximizer. Being an REU maximizer, he reflectively endorses his decision to violate Independence in virtue of his sensitivity to global properties.

This approach is implausible for two reasons. The first is that if REU maximizers were common enough to explain the experimental results that support prospect theory, then we should see the impact of their disposition more broadly in society—as in the Primary Elections case, for example. Whatever tendency explains the disposition to violate Independence, it seems more context sensitive than REU.

The second reason this response is implausible is that given the foundational differences in mental constitution that my theory suggests there are between REU and EUM agents, it isn’t clear how both REU and EUM agents would come to exist in the population. That is, it would be odd if people were just randomly constructed with desires that recommended one kind of ‘ought’ judgment or the other. You might think it should have some kind of genetic trigger or run in families. But we have no reason to think that is true.

6.8.4 Failed Reply 3: Practical Disjunctivism

The objection from reflective irrationality assumes that the mental states that motivate predictably irrational behavior do so by way of qualitatively similar teleological functionalist aim properties. The similarity between these properties in rational mental states and irrational mental states is so high that people lose the ability to distinguish between them. This creates a problem for TF. But it is conceivable that people do not lose the ability to distinguish between irrational and rational aims because they are qualitatively identical. Rather, people might suffer from an inability to distinguish a real qualitative difference. If there is a real qualitative difference, and some people just cannot detect it, then it is possible that only rational mental states have the right qualitative property and reflective irrationality is a problem of misidentifying irrational aim properties as rational ones.\[18\]

\[18\]This does involve the possibility of misidentifying a first-personal property. Some theories reject this possibility. But Disjunctivism, the theory that grounds this type of response, rejects that idea that we have perfect epistemic access to first-personal qualities.
This is the practical analog of a theory in the philosophy of perception called *Disjunctivism*. The idea there is to explain hallucinations not in terms of qualitatively identical first-personal experiences, but rather to suggest that a hallucination is a case of an agent temporarily losing their ability to distinguish two qualitatively different experiences.

While this approach does provide for the logical possibility of a TF explanation of reflective irrationality, it will be at least as controversial as its cousin in the philosophy of perception—which is to say very much so. An additional problem with this approach is that it requires that people who endorse alternatives to EUM are under a kind of practical analog of a sustained hallucination. Authors like McClennen and Buchak do not reflectively endorse Independence violations momentarily. Rather, they have long held opinions about the issue. So if we were to take this approach, it would suggest that there is some kind of sustained introspective blindness occurring here. In addition to being a strong claim, there doesn’t seem to be anything the reflectively irrational could do to escape their situation—no course of reasoning they could take to recognize the appeal of rationality. In this sense, this answer fails the Deontic Appeal and Extensional Adequacy success conditions for theories of practical rationality. The goal, we should hope, is to provide a story that could conceivably lead to convincing REU theorists one day. But this answer eliminates that possibility.

But, what is more telling from my own point of view is that ‘irrational’ aims do, introspectively, seem like aims to me as well, even though I endorse EUM. So according to the TF story, they are for the purpose of getting me to do something. This is as it should be. When I have a desire that I do not endorse—that does not create in me a judgment that it would be a good thing to do—that desire is still very much for the purpose of getting me to form that judgment. In a case of weakness of will, I might act to satisfy the desire without the corresponding judgment that I ought to. So having non-EUM attitudes that *seem like aims* cannot be enough to separate those that endorse EUM from those who don’t.

---

19 Notice that on TF it is entirely possible as a causal matter that this happens. The causal powers of desire might come apart from the expression of the Teleological Functional role, and might, plausibly, execute the parts of that role inconsistently.

20 One might be inclined to think that it is the presence of the judgment that I ought to satisfy a desire that gives the desire deontic normative force. But this wouldn’t be enough to provide an error theory for reflective irrationality because the REU agent does have the judgment that they ought to violate Independence.
6.8.5 A Working Reply: Desires for Risk Profiles

A risk profile is facing the prospect of an outcome or set of outcomes at a given probability. There is a very subtle psychological difference between treating risk as important for decisions and treating a risk profile as an object of desire. Theorists who violate Independence might be making the mistake with two parts: thinking of risk as an object of desire and then constructing a model of rationality as if that desire was a feature of rationality itself.

An agent treats a risk profile as an object of desire when he thinks of it as attractive or unattractive independently of its attachment to particular outcomes. An REU agent, for example, determines his risk function by thinking about probabilities, and the weight that outcomes get at particular probabilities based on their ranking. This happens regardless of what the outcomes themselves actually are. So this is a mental state that is different from an attitude towards apples, or money, or whatever the agent might get at the end of the lottery. I claim that this attitude towards risk profiles themselves is another desire. Because it is an additional desire, the correct way to represent it is by redifferentiating outcomes to include the fact that they came about as the result of a particular preferred lottery profile. A redifferentiation like this would make the agent representable by EUM.

Having a desire for a risk profile does not necessarily imply that an agent will regret his choice afterwards, or pay to reverse his decision if possible. Take the case of Jeff from the beginning of this chapter. If Jeff has a desire for risk profiles, then he doesn’t necessarily need to be willing to pay to trade nothing for nothing when he could have had gloves. Because a desire for a risk profile is a kind of sensitivity to a history, or path to an outcome, and paying to reverse a decision changes the history, it isn’t clear that paying to reverse a decision would satisfy a desire like that.

Forming a desire for a risk profile is an easy thing to do and can easily go unnoticed. It is also something that is more likely if you think about a decision too hard. A person might think, “How good are these outcomes?” And then, consciously recognizing that risk should be factored into decision making think “What are the probabilities?” and “How should I take the risk into account?” This matches up with the three questions Buchak argues rational agents need to answer. But what the agent doesn’t realize is that his answer to the first question included an answer to the third question; it just presented that answer in a different form. The moment the agent asks that third question, he might start to form another desire—a desire for a risk profile. This desire was formed as a consequence of attending to risk itself and placing variables representing outcome rankings in place of possible states of the world. A person might develop desires reflecting their attitudes about this abstracted scenario and then then weight that desire against desires for states of the
world themselves. Because a risk profile desire is created when attention is focused on an object that abstracts away from details of the actual decision, it will plausibly conflict to some degree with desires for the original outcomes in question.

There is nothing wrong with having a desire for a risk profile in itself. A desire for a risk profile can weigh against one’s desire for original outcomes. But a desire for a risk profile is just a contingent desire that someone may or may not have and that is inessential to practical reasoning. As such, global sensitivity is not something that should be represented in the structure of the representation itself. Rather, it is just a component of the EUM preferences of certain individuals.21

What happens in cases like this is that the risk gets double counted—once in assessing how much I desire an outcome, and then once again in an evaluation of a risk-profile. The difference between thinking in terms of risk-profiles, and being attracted to efficacy at an outcome is subtle enough that a theorist could consistently fail to appreciate it.

6.8.6 A Working Reply: Evolved Heuristics, the Discovery of Risk, and Extrapolating Intuitions to Principles

Thousands of years ago, in whatever context human beings evolved the capacity to deliberate about what they ought to do, they didn’t deliberate using judgments about probability. The beginnings of the application of probability didn’t occur until the 17th century (Fermat and Pascal 1654). So whatever ancient deliberators were doing, it wasn’t comparing values for their \( v(\phi) \) function.

The same is probably true for children. Children might have a basic understanding of how the likelihood of the success of an action impacts its attractiveness as a means, and yet they do not know the basics of probability theory until they are taught.

This implies that the notion of goodness that agents naturally use isn’t quite the same as the notion of utility. Utility numbers are generated by way of an artificial scale—a scale that uses probabilities. So the notion of ‘twice as good as’ that corresponds to utility can’t be the notion that people before the 17th century used, or that a child might use, in uttering that phrase.22

21 People with desires for risk profiles are still EUM because if they didn’t have those desires they would display more straightforward EUM behavior. If you know the probabilities involved, you can satisfy a desire for a risk profile with a probability of 1—you take a sure shot at facing a particular risk profile.

22 A similar distinction arises with respect to statements about how good objects are rather than how good outcomes are. Strictly speaking, a utility attaches to a state of the world—e.g., getting
But children, and adults before the 17th century, still come to conclusions about what to do. So whatever system is present in our psychologies before training, it is capable of making actions seem like the right thing to do even without plugging in information about probabilities. An easy error to make that could explain reflective irrationality is if theorists extrapolated from their untrained natural choice intuitions to conclusions about the nature of decision under probabilistically specified risk. If the tendency to violate Independence is the expression of beliefs and desires before probability theory is fully taken into account, then this is a way Teleological Functionalism could work given reflective irrationality. TF continues to work because there is only one set of states with first-personal aim properties—the states I specified in chapter 2. These states are naturally vague, but they are about a specific subject matter and ought to become more precise as we gain knowledge about their referent.

We can get clearer about this idea by looking at what happens to deliberation when someone learns about probability theory. Suppose that we are born with a capacity to make vague judgments about likelihood without probability. Likelihood has a particular subject matter, but it is something that we have yet to specify exactly. So when an agent learns about probability, she learns something that is relevant to the deliberative process she had been using all along. She learns that the vague notion of likelihood is actually trying to track something which can be specified in a more precise way. She learns that lotteries have specifiable parts that fit together in regularly comparable ways, for example. Being tied to her notion of likelihood, her basic ‘ought’ judgments become more precise as well. So there is a sense in which the TF mental states I have could naturally aim at something non-EUM and yet it could be appropriate to revise them into EUM compliance. Specifically, in being about a mind independent notion of likelihood, or efficacy, that can get revised with more information.

This creates room for the possibility that the actions we find intuitively correct might be different from the actions our desires recommend when we are precise about an apple, not the apple itself. But we talk in terms of the value of objects all the time. A gold bar might be worth twice as much as a silver one. The language of how good something is that applies in EUM might be unnatural for people who tend to think and talk in terms of how good objects are. This could add another complication to the story presented in this error theory, but I’ll leave it out of the main text in order to highlight the main idea.

23One benefit of the informal presentation of the content of our basic ‘ought’ judgments in chapter 3 is that it is more broad than the content expressed in the \( v(\phi) \) function presentation. That is, judgments about likelihood and efficacy can occur in contexts of uncertainty as well as contexts of risk. The \( v(\phi) \) function represents the implications of desires in contexts of risk, but ancient man didn’t deliberate about risk; at best he deliberated about uncertainty.
the judgments they require. The same system, with the same Teleological Functional
roles, has the capacity to make vague recommendations about what you ought to do
based on likelihood and desire. Operating with these vague inputs, the system might
have a tendency to exaggerate the disvalue of possible loss or value of a possible gain,
leading to Prospect Theory like intuitions. But when we discover that these mental
states are about something that we can be more precise about, then we are forced
to rein in our intuitions about these cases and be EUM.

Notice that this kind of exaggeration is especially plausible because there is no
scale of goodness corresponding to utility until probability is introduced. So suppose
an agent wanted to compare the value of an equal likelihood of \( \omega_1 \) with an equal
likelihood of \( \omega_2 \). That the shot at \( \omega_1 \) has any fixed value relative to the shot at
\( \omega_2 \) is not a well defined idea because we do not have any conceptual analog of a
utility function or \( v(\phi) \) function to explain it. Before probability theory, there is
nothing forcing consistent treatment of risk. So intuitions about what one ought
to do might develop differently from what the content of the relevant judgments
actually recommends.

If a theorist starts with his intuitions about individual decisions and then extrap-
olates to a theory of rationality generally, he might make the mistake of extrapolating
on the basis of a system of heuristics. Assuming that the theory I presented in the
first three chapters is correct, and that we are indeed explaining an error, this mistake
can be avoided by looking at the intended subject matter of judgments about the
relationship between likelihood and value. This is also why checking for the impact
of Independence violation on peripheral concepts and practices can be a more effec-
tive method of revealing the nature of rationality than looking directly at whether
an action is intuitive for you. We are not checking for coherence with other con-
cepts. Rather, we are looking at what the effective expression of the \( TF \) roles of our
mental states when the risk is precisely specified. Looking at the strange peripheral
consequences of Independence violation helps separate our deliberation from evolved
heuristics and allows us to focus on what that deliberation is actually about.

6.9 Conclusions

6.9.1 On This Chapter

Given the considerations that Buchak takes to weigh on the decision between EUM
and REU, REU does seem like the more attractive option. However, I’ve brought
up a variety of new considerations that might convince someone that REU is both
unlikely and the result of understandable mistakes in the evaluation of the nature of
reasoning.

One thing I have not done in this chapter is argue against a position where Independence violation is more context specific than REU. Considerations like the antibiotic case or the primary elections case seem to suggest that considerations that make Independence violation rational, if there are any, apply in a more context sensitive way than REU implies. But the examples I raise in this chapter do not address this possibility, they just bring up cases where Independence does seem to apply.

I do not take myself to have provided an argument against this position. In the presence of a well-developed theory, an approach like this might be viable. One thing to note is that the more context sensitive Independence violation is meant to be, the more plausible it is that it is appropriately captured by redifferentiating outcomes to include a desire for the property justifying the purported violation. For example, if global properties provided insurance in a particular case, then this insurance should add a premium to the utility of an outcome. REU avoids this problem by being the result of an attitude that tracks risk itself, and, as such, does not plausibly bring up new reasons for a choice while allowing for the possibility of changing the weight of the reasons an agent already faces. If a new reason for action arises in a specific context, then that reason is plausibly a component of an outcome. As such it is something EUM might capture.

6.9.2 Regarding the Overall Project

We set out with a goal of providing a plausible quasi-Humean interpretation of EUM that captured a common sense notion of what EUM was about. Even if the view doesn’t convince reflective Independence violators to change sides, it does demonstrate the possibility of a common sense EUM interpretation despite concerns in this vein. This interpretation also provides an explanation of what a sufficiently broad set of agents might find unappealing about violating EUM, which is the challenge that Sen started out with.

A consequence of the internalist position I presented is that there is no deontic normative view from nowhere. Exploitability, for example, is not bad in itself—its badness consists in its being incompatible with an agent’s aims from his own point of view. For this reason, I am pessimistic about the prospects for any argument about the nature of rationality to be purely conceptual. Rather, we have to give a description of functional interrelations between mental states, and specifications of concepts employed, that appear, through observation and reflection, to match what actual human beings seem to aspire to.
We can see the manifestation of this in a traditional approach to arguing that a position is irrational. Exploitability itself, for example, is not the problem. The fully described problem is that exploitability seems unattractive from our point of view, so it is implausible that we are the kinds of beings who aspire to behavior that includes exploitability. This places arguments from exploitability within the same genus as arguments about the strangeness of the concepts of efficacy they employ or the mental states they require. In each case, it is the distance between the picture those theories present and our common understanding of ourselves, and not some essential or metaphysical property of normativity as such, that makes these theories implausible. There could, conceivably, be REU agents out there. My claim is that they are not us.

My final general conclusion is that it seems important that a decision theory present a picture of practical rationality that actual people could live up to some of the time. This means that the reasoning process that they employ to make decisions between basic lotteries cannot be so complicated as to rule out all but expert decision theorists. So while there may be multiple axiomatizations of Expected Utility Maximization, it seems reasonable to think that agents ought to be able to make rational choices without necessarily being able to run through the proof stemming from some of these more complex pictures. Even though the concepts of efficacy and NIE are not themselves very far removed from the idea of expected utility maximization in itself, they have the added benefit that most intelligent people could approximate EUM if they had access to the concepts that I take here to be most basic. I would not expect the average person to be able to conclude that he ought to be EUM from the axioms of Von Neumann and Morgenstern. This is just another reason to think that these axioms are not the concepts that people employ when they are reasoning about what they ought to do.

While people do systematically deviate from EUM, and while some theorists endorse those deviations as rational, Teleological Functionalism substantially restricts our options for explaining this deviation. Broadly speaking, it rules out explanations where there is another system of mental states that operate alongside beliefs and desires and have their own teleological aim properties. In cases like this, Teleological Functionalism would return the result that these other systems are just as rational as action in compliance with EUM. Teleological Functionalism forces an explanation of reflective irrationality whereby agents and theorists are making mistakes about the expression of beliefs and desires as a result of challenges that naturally arise from the instantiation of these mental states in human beings. None of this is to give a complete error theory of reflective irrationality, however. If, upon investigation, it turned out that Independence violation was a response to an entirely different set of
mental states, then that would be a problem for the theory presented here.
Bibliography


Chapter 7
Appendix

7.1 Introduction

Definitions:

- a set $\Omega$ of outcomes that are the possible consequences of actions. We will assume that the set $\Omega$ is finite and we will index the members by $n = 1, ..., N$.

- a set $L$ of possible lotteries, or lists $(l_1 = (p_1, ..., p_N))$, that assign a probability $p_n \in [0, 1]$ to each member of $\Omega$. $\sum_{n=1}^{N} p_n = 1$

- a partition of a lottery $\Phi_l$ is a set whose members, $\phi_{li} \in \Phi_l$ are lists of changes in probability associated with outcomes such that $\sum_{i=1}^{N} (p_{l\phi_{li}}, ..., p_{N\phi_{li}}) = (p_1, ..., p_N)$. In other words, if you sum the all the changes in probability associated by members of the partition with each outcome, you will get the probability $l$ assigns to that outcome.

1. Example: $l_1 = (p, \omega_1; (1 - p), \omega_2)$. You can construct partitions by dividing a lottery into equal parts: $\phi_{l11} = (p/2, \omega_1; (1 - p)/2, \omega_2)$ and $\phi_{l12} = (p/2, \omega_1; (1 - p)/2, \omega_2)$. You can construct partitions by assigning zero change in probability to one outcome in some members of the partition: $\phi_{l11} = (0, \omega_1; (1 - p), \omega_2)$ and $\phi_{l12} = (p, \omega_1; 0, \omega_2)$. You can also assign negative changes in probability in members of a partition: $\phi_{l11} = (0, \omega_1; 0, \omega_2), \phi_{l12} = ((p - 1), \omega_1; (1 - p), \omega_2)$ and $\phi_{l13} = (1, \omega_1; 0, \omega_2)$. (See Chapter 3).
2. A partition $\Phi_2$ can also be made of a member of another partition $\phi_{l1}$ such that $\sum_1^N (p_{1,\phi_{2i}}, \ldots; p_{N,\phi_{2i}}) = (p_{1,\phi_{l1}}, \ldots; p_{N,\phi_{l1}})$.

3. For ease of notation, let $(r, \phi)$ represent the member of some partition that obtains from distributing $r$ across the probabilities $\phi$ assigns to outcomes. So if $\phi = (p_1, \omega_1; \ldots; p_n, \omega_n)$ then $(r, \phi) = (rp_1, \omega_1; \ldots; rp_n, \omega_n)$.

- a function $v(\phi) \to \mathbb{R}$ which represents the instrumental value of a member of a partition, $\phi$, in any lottery in which it is embedded.
  1. $v(l_1) = v(\phi_{l1})$ where $\phi_{l1}$ is the only member in the partition of $l_1$.
  
- a null member, $\phi_{NULL}$, of a partition is a member such that for all $\phi$, $v(\phi_{NULL}) = v(0, \phi) = 0$.

7.2 Chapter 3

The value function $v(\phi)$ has three parts

- **Definition of Preference:** For any two lotteries, $l_1$ and $l_2$, $l_1 \succeq l_2$ if and only if $v(l_1) \geq v(l_2)$, where $v(l_1) = v(\phi_{l1})$ when $\phi_{l1}$ is the only member in the partition of $l_1$.

- **No Interaction Effects:** for any set of outcomes, $\omega_A \ldots \omega_N$, and corresponding changes in probability, $p_1\ldots p_N$, $v(p_1, \omega_A; \ldots p_N, \omega_N) = v(p_1, \omega_A) + \ldots + v(p_N, \omega_N)$.
  1. This holds for all $\omega_A \ldots \omega_N$, including when outcomes are repeated—such as when $\omega_A = \omega_N$. So, for example, for any $p$ and $\omega_A$, $v(p, \omega_A) = x(p/x, \omega_A)$ by NIE.

- $v(\phi_{NULL}) = v(0, \phi) = 0$.

**The Definition of a Fixed Tradeoff Rate** Let $(p, \omega)$ stand for the member of a partition that obtains from setting a $p$ change in probability to outcome $\omega$ and a zero change in probability to all other outcomes. A fixed tradeoff rate, $r/s$, exists

1I take it that there is no significant difference between thinking of lotteries as lists of probabilities and partition members being lists of “changes in probability” in the institution of this convention. If he or she prefers, the reader is free to think of lotteries here defined as changes in probability associated with outcomes relative to zero probability of those outcomes. I do not believe this change would alter the substantive content of anything in this project.
between \( \omega_A \) and \( \omega_B \) just in case, \((s, \omega_A)\) makes the same contribution to the value of a gamble as \((r, \omega_B)\).

The following three statements capture the formal definition of the Fixed Tradeoff Rate, \(r/s\) between \( \omega_A \) and \( \omega_B \). For all \( \omega_A, \omega_B, p, r \) and \( s \):

1. If \((p, \omega_B) \neq \phi_{\text{NULL}}\), then \(v(p, \omega_A) = (r/s)v(p, \omega_B)\).
2. If, for any nonzero \( p \), \( v(p, \omega_A) \neq 0 \) and \( v(p, \omega_B) \neq 0 \), then \(v(s, \omega_A) = v(r, \omega_B)\).
3. \(v(p, \omega_A) = (r/s)v(p, \omega_B)\), then \(v(p, \omega_A; t, \omega_B) = v(rp/s + t, \omega_B)\).

7.2.1 Theorem 1: Linearity

For all \( x, p, \) and \( \phi \), \( xv(p, \phi) = v(xp, \phi)\)\(^2\).

Proof:

Let \( \phi = (r_1, \omega_1; ...; r_N, \omega_N) \).

\(v(xp, \phi) = v(xpr_1, \omega_1; ...; xpr_N, \omega_N)\).

\(v(xp, \phi) = v(xpr_1, \omega_1) + ... + v(xpr_N, \omega_N)\) by NIE.

Any \(v(xpr, \omega_A) = xv(pr, \omega_A)\) by NIE.

\(v(xp, \phi) = x[v(pr_1, \omega_1) + ... + v(pr_N, \omega_N)]\) by substitution.

\(v(p, \phi) = v(pr_1, \omega_1) + ... + v(pr_N, \omega_N)\) by NIE.

\(v(xp, \phi) = xv(p, \phi)\) by substitution.

7.2.2 Theorem 2.1: That for any outcomes, \(\omega_A\) and \(\omega_B\), that \(v(\phi)\) is in compliance with the above axioms implies that a tradeoff rate ((1)-(3) for some \(r/s\)) exists between \(\omega_A\) and \(\omega_B\).

Proof:

Suppose that at probability \( t \), \(v(t, \omega_B) \neq 0\).

Because \(v(t, \omega_A)\) and \(v(t, \omega_B)\) are non-zero real numbers, there exists an \( x \) such that

\(v(t, \omega_A) = xv(t, \omega_B)\).

\(tv(1, \omega_A) = txv(1, \omega_B)\) by linearity.

\(v(1, \omega_A) = xv(1, \omega_B)\).

\(^2\)Linearity could also be written \(xv(\phi) = v(x, \phi)\) for the special case where \(p = 1\) and all the same results would follow.
\( pv(1, \omega_A) = pxv(1, \omega_B) \) for all \( p \).

\( v(p, \omega_A) = xv(p, \omega_B) \) for all \( p \), by linearity.

(1) Let \( r/s = x \). Then \( v(p, \omega_A) = (r/s)v(p, \omega_B) \).

\( sv(p, \omega_A) = rv(p, \omega_B) \).

(2) So \( v(s, \omega_A) = v(r, \omega_B) \) by linearity.

Suppose (1): \( v(p, \omega_A) = (r/s)v(p, \omega_B) \).

\( v(p, \omega_A; t, \omega_B) = v(p, \omega_A) + v(t, \omega_B) \) by NIE.

\( v(p, \omega_A; t, \omega_B) = v(p, \omega_A) + (s/r)v(t, \omega_B) \) by (1).

(3) \( v(p, \omega_A; t, \omega_B) = v(p + st/r, \omega_A) \) by linearity and NIE.

7.2.3 Theorem 2.2: That the tradeoff rate \( r/s \) between \( \omega_A \) and \( \omega_B \) is fixed. That is, if there exists a tradeoff rate \( x/y \) between \( \omega_A \) and \( \omega_B \) then \( x/y = r/s \).

Proof:

Suppose that for some change in probability \( t \), \( x/y \) is the tradeoff rate between \( \omega_A \) and \( \omega_B \).

\( v(t, \omega_A) = (t/p)v(p, \omega_A) \) by Linearity.

\( (t/p)v(p, \omega_A) = (x/y)(t/p)v(p, \omega_B) \) substituting equivalent terms.

\( v(p, \omega_A) = (x/y)v(p, \omega_B) \).

So \( x/y = r/s \) by Theorem 2.1.

7.2.4 Theorem 3: That if \( l_1 \succ l_2 \) then there exist two partitions \( \Phi_{l_1} \) and \( \Phi_{l_2} \), each with \( n \) members, such that for all \( i \) from 1 to \( n \), \( v(\phi_{l_1i}) > v(\phi_{l_2i}) \).

Proof:

\( v(l_1) > v(l_2) \) by the Definition of Preference

\( nv(1/n, l_1) > nv(1/n, l_2) \) by Linearity

\( v(1/n, l_1) > v(1/n, l_2) \)

Each member of the relevant partitions is just the original lottery with the probabilities associated with outcomes multiplied by \( 1/n \).
Continuity - for any three lotteries \( l_1, l_2, l_3 \in L \), if \( l_1 \succeq l_2 \succeq l_3 \), then there exists a \( t \in [0, 1] \) such that the agent is indifferent between a \( t \) probability of \( l_1 \) coupled with a \((1 - t)\) probability of \( l_3 \) and getting \( l_2 \) with probability 1.

Let \((p_1...p_N), (r_1...r_N)\), and \((s_1...s_N)\) be the probabilities that \( l_1, l_2, \) and \( l_3 \) respectively to the outcomes \((\omega_1...\omega_N)\).

**Proof:**

\[
v(l_1) \geq v(l_2) \geq v(l_3)
\]

by the Definition of Preference.

Because \( v(\phi) \) maps to the real numbers, there exists a \( t \) such that

\[
tv(p_1, \omega_1; ...; p_N, \omega_N) + (1 - t)v(s_1, \omega_1; ...; s_N, \omega_N) = v(l_2).
\]

\[
v(tp_1 + (1 - t)s_1, \omega_1; ...; tp_N + (1 - t)s_N, \omega_N) = v(l_2)
\]

by Linearity and No Interaction Effects.

\[
(t, l_1; 1 - t, l_3) \sim l_2
\]

by the Definition of Preference.

### 7.2.6 Independence from the Definition of \( v(\phi) \)

**Independence:** If \( l_1 \succeq l_2 \) then \((t, l_1; 1 - t, l_3) \succeq (t, l_2; 1 - t, l_3)\).

Let \((p_1...p_N), (r_1...r_N)\), and \((s_1...s_N)\) be the probabilities that \( l_1, l_2, \) and \( l_3 \) respectively to the outcomes \((\omega_1...\omega_N)\).

**Proof:**

\[
l_1 \succeq l_2.
\]

\[
v(l_1) \geq v(l_2)
\]

by the Definition of Preference.

\[
v(p_1, \omega_1; ...; p_N, \omega_N) \geq v(r_1, \omega_1; ...; r_N, \omega_N).
\]

\[
tv(p_1, \omega_1; ...; p_N, \omega_N) \geq tv(r_1, \omega_1; ...; r_N, \omega_N).
\]

\[
vtv(p_1, \omega_1; ...; p_N, \omega_N) + (1 - t)v(s_1, \omega_1; ...; s_N, \omega_N) \geq
\]

\[
vtv(r_1, \omega_1; ...; r_N, \omega_N) + (1 - t)v(s_1, \omega_1; ...; s_N, \omega_N).
\]

\[
vtv(tp_1 + (1 - t)s_1, \omega_1; ...; tp_N + (1 - t)s_N, \omega_N) \geq v(tp_1 + (1 - t)s_1, \omega_1; ...; tp_N + (1 - t)s_N, \omega_N)
\]

by Linearity and No Interaction Effects.

\[
(t, l_1; 1 - t, l_3) \preceq (t, l_2; 1 - t, l_3)
\]

rearranging terms, and by the definition of preference.
7.3 Chapter 5

7.3.1 Mission to Mars

Problem Specifications

- Blow up: $\omega_1$
- Stranded: $\omega_2$
- Arrive with backup: $\omega_3$
- Arrive: $\omega_4$
- $\omega_4 >_V \omega_3 >_V \omega_2 >_V \omega_1$
- Backup Rocket A: $q(\omega_3) + (1-q)(\omega_1)$
- Backup Rocket B: $s(\omega_3) + (1-s)(\omega_2)$
- Overall Strategy A: $p(\text{arrive}) + (1-p)(\text{backup rocket A})$
  \[ = p(\omega_4) + (1-p)(q)(\omega_3) + (1-p)(1-q)(\omega_1) \]
- Overall Strategy B: $p(\text{arrive}) + (1-p)(\text{backup rocket B})$
  \[ = p(\omega_4) + (1-p)(s)(\omega_3) + (1-p)(1-s)(\omega_2) \]
- $u(\omega_1) = 1$
- $u(\omega_2) = 2$
- $u(\omega_3) = 3$
- $u(\omega_4) = 4$
- $p = .9$
- $q = .99$
- $s = .8$
Risk-Weighted Utility Functions

\[
REU_A = u(\omega_1) + r(q)(u(\omega_3) - u(\omega_1))
\]

\[
REU_B = u(\omega_2) + r(s)(u(\omega_3) - u(\omega_2))
\]

\[
REU_{OA} = u(\omega_1) + r(q + p - pq)[u(\omega_3) - u(\omega_1)] + r(p)[u(\omega_4) - u(\omega_3)]
\]

\[
REU_{OB} = u(\omega_2) + r(s + p - ps)[u(\omega_3) - u(\omega_2)] + r(p)[u(\omega_4) - u(\omega_3)]
\]

With numbers.

\[
REU_A = 1 + r(.99)2
\]

\[
REU_B = 2 + r(.8)
\]

\[
REU_{OA} = 1 + r(.99 + .9 - .9(.99))2 + r(.9)
\]

\[
REU_{OB} = 2 + r(.8 + .9 - .9(.8)) + r(.9)
\]

Conditions for Independence Failure

1. If \(REU_A > REU_B\) then \(2r(.99) - r(.8) > 1\)

2. If \(REU_{OB} > REU_{OA}\) then \(2r(.999) - r(.98) < 1\)

Both of these conditions can hold while \(r()\) is continuous and strictly increasing \((r() \text{ increases more slowly above } p = .99)\). For example, suppose \(r(.99) = .51\), \(r(.8) = .01\), \(r(.999) = .52\), and \(r(.98) = .5\).

7.3.2 Antibiotics

Problem Specifications

- infection gets worse: \(\omega_1\)
- no effect: \(\omega_2\)
- infection improves: \(\omega_3\)
- Antibiotic A: \(p(\omega_3) + s(\omega_2) + (1 - p - s)(\omega_1)\)
- Antibiotic B: \(t(\omega_3) + (1 - t)(\omega_2)\)
- Placebo: \(\omega_2\)
• Trial A: \( q(antibiotic\ A) + (1-q)(placebo) \)
  \[= qp(\omega_3) + (qs + (1 - q))(\omega_2) + q(1 - p - s)(\omega_1) \]
• Trial B: \( q(antibiotic\ B) + (1-q)(placebo) \)
  \[= qt(\omega_3) + (q(1 - t) + (1 - q))(\omega_2) \]
• \( u(\omega_1) = 1 \)
• \( u(\omega_2) = 3 \)
• \( u(\omega_3) = 5 \)
• \( p = .9 \)
• \( s = .05 \)
• \( t = .8 \)
• \( q = .75 \)

**Risk-Weighted Expected Utility Functions**

\[
REU_A = u(\omega_1) + r(p + s)(u(\omega_2) - u(\omega_1)) + r(p)(u(\omega_3) - u(\omega_2))\\
REU_B = u(\omega_2) + r(t)(u(\omega_3) - u(\omega_2))\\
REU_P = u(\omega_2)\\
REU_{TA} = u(\omega_1) + r(1 - v(1 - p - s))(u(\omega_2) - u(\omega_1)) + r(vp)(u(\omega_3) - u(\omega_2))\\
REU_{TB} = u(\omega_2) + r(vt)(u(\omega_3) - u(\omega_2))
\]

With numbers.

\[
REU_A = 1 + r(.95)(2) + r(.9)(2)\\
REU_B = 3 + r(.8)(2)\\
REU_P = 3\\
REU_{TA} = 1 + r(1 - .75(1 - .9 - .05))(2) + r(.9 * .75)(2)\\
REU_{TB} = 3 + r(.8 * .75)(2)
\]

**Conditions Required for Independence Failure**

1. \( REU_A > REU_B \rightarrow 2r(.95) + 2r(.9) > 2 + 2r(.8) \)
2. \( REU_{TB} > REU_{TA} \rightarrow 2 + 2r(.6) > 2r(.9625) + 2r(.675) \)

This is possible assuming \( r() \) is continuous and strictly increasing (assume that when \( p > .8, r(p) \cong .5 \) and when \( p < .8, r(p) \cong 0 \)).
7.3.3 Riflery Competition

Problem Specifications

- Bronze Medal: $\omega_1$
- Silver Medal: $\omega_2$
- Gold Medal: $\omega_3$
- Rifle 1: $p\omega_3 + (1 - p)\omega_1$
- Rifle 2: $\omega_2$

- Suppose that the value of medals is independent. So $REU(2\text{ medals}) = REU(\text{medal 1}) + REU(\text{medal 2})$.

- $t =$ the number of tournaments
- $u(\omega_1) = 1$
- $u(\omega_2) = 2$
- $u(\omega_3) = 4$
- $t = 100$
- $p = .6$

Risk-Weighted Expected Utility Functions

\[
REU_{tR1} \cong .6tu(\omega_3) + .4tu(\omega_1)
\]
\[
REU_{tR2} \cong tu(\omega_2)
\]
\[
REU_{R1} \cong u(\omega_1) + r(p)(u(\omega_3) - u(\omega_1))
\]
\[
REU_{R2} \cong u(\omega_2)
\]

With numbers.

\[
REU_{tR1} \cong 280
\]
\[
REU_{tR2} \cong 200
\]
\[
REU_{R1} \cong 1 + r(p)(3)
\]
\[
REU_{R2} \cong 2
\]

Conditions for Sequential Inconsistency
1. \( r(.6) < .333 \)

So as long as \( r(.6) < .333 \) choosing Rifle 1 will seem preferable in each individual instance, even though the agent would prefer to end up with the expected mix of medals rather than 100 silver medals.