Non-Ideal Decision Theory

by

Sven Moritz Silvester Neth

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Philosophy

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lara Buchak, Co-Chair
Professor John MacFarlane, Co-Chair
Professor Wesley Holliday
Professor Amy Rose Deal

Fall 2023

Non-Ideal Decision Theory

Copyright

Sven Moritz Silvester Neth

2023

Abstract

Non-Ideal Decision Theory

by

Sven Moritz Silvester Neth

Doctor of Philosophy in Philosophy

University of California, Berkeley

Professor Lara Buchak, Co-Chair

Professor John MacFarlane, Co-Chair

Standard decision theory has some striking consequences. First, if you have any irrational preferences, it does not make sense to ascribe credences to you—decision theory treats you as having no opinions at all. Second, you should always prefer to look at more information before making a decision if the information is free, even if you think the information is likely biased.

These consequences seem wrong. Your preferences are sometimes irrational, but you clearly still have opinions about things. If we could only ascribe credences to agents with perfectly rational preferences, decision theory couldn't give any useful advice to imperfect agents like us. Furthermore, people sometimes prefer *not* to look at more information before making a decision. For example, when grading a paper, I prefer not to know the student's name. This behavior seems perfectly rational, even laudable.

Does this mean that decision theory is broken? No. I argue that we can fix its problems. The key to ascribing credences to non-ideal agents is to start by looking at *comparative probability judgments*, like thinking it's equally likely to be sunny and rainy tomorrow. Comparative probability is tied to preferences in a straightforward way—you think that sunshine and rain are equally likely if you are indifferent between betting on them. If sunshine and rain are the only two possibilities for the weather tomorrow and you think they're equally likely, you assign probability .5 to both. I show that if your

preferences satisfy some minimal constraints, we can extend this procedure to fix your entire credence function while allowing many of your preferences to be irrational (Chapter 2).

The key to explaining why it can be rational to reject free information is to recognize that we can be *uncertain about how we will react to evidence*. If I were certain that I always respond to evidence in a perfectly rational way (by 'conditionalization'), then perhaps it would be rational to look at my students' names before grading papers. Indeed, certainty that one will react to evidence in a perfectly rational way is an assumption built into Good's well-known theorem about the value of information. But this assumption might fail. I know that I'm not always rational—for example, I might give too much weight to the fact that George got an 'A' on the first paper and treat this as better evidence than it is that his current paper deserves a high mark. Once I take this possibility into account, I might be better off ignorant (Chapter 3).

Uncertainty about how we will react to evidence has other consequences as well. For example, it can lead us to make sequences of choices which, taken together, yield sure loss. Many decision theorists have claimed that such choices always indicate irrationality. I argue that this bit of conventional wisdom needs revision (Chapter 4).

One upshot is that by widening the scope of decision theory to include non-ideal agents, we enable the decision theorist to give vindicating explanations of common phenomena, like having opinions without being fully rational and avoiding information before making a decision. Another upshot is that a more sophisticated decision theory is relevant for designing beneficial AI systems, since existing ideas for doing so assume an implausibly strong conception of rationality (Chapter 5).

Das Hauptproblem aller nur rationalen Weltzugänge: man kriegt zu wenige Aspekte gleichzeitig zu fassen, und dieses Orientierende, der gefühlsmäßige Überblick, die Ansicht des Ganzen in ihrer Irgendwiehaftigkeit kann durch keine Schärfe der Analyse im Einzelnen usw.

Rainald Goetz, *loslabern.*

Acknowledgments

First, I would like to thank my co-chairs Lara Buchak and John MacFarlane for many rounds of incredibly helpful feedback on my writing and for teaching me how to do good philosophy. I would also like to thank Wesley Holliday for serving on my committee and helping to sharpen my arguments. Beyond my committee, I would like to thank Barry Stroud. Without taking Barry's class on Theory of Knowledge in Fall 2014 and learning about grue, I'm not sure I would have decided to do philosophy.

At Berkeley and beyond, I have found an amazing philosophical community. In particular, I would like to thank Mathias Böhm, Mikayla Kelley and Adrian Ommundsen for their friendship. Mathias and I started graduate school together and have talked about everything in this dissertation. I met Mikayla when she was studying math at Berkeley and transferring to philosophy. Since then, she has been a great friend and now also co-author. I met Adrian last year when he was visiting Berkeley and since then we've spent much time talking philosophy.

More generally, thanks to Berkeley's awesome community. Many times when I felt confused and stuck, going to 301 kept me on track. Thanks to Alexander Kocurek ('Arc'), Ethan Jerzak, Emily Perry and James Walsh for hanging out. I've spend many happy hours at the Wollheims. I'm thankful for many lit parties. I'm also thankful to everyone who talked to me about grue, even though it didn't end up in this dissertation. You helped me to stay excited about philosophy.

For helpful comments and discussions on the material of this dissertation, I would also like to thank Morgan Rachel Connolly, Reid Dale, Shamik Dasgupta, Yifeng Ding, Kevin Dorst, Kenny Easwaran, James Evershed, Johann Frick, Aglaia von Götz, Anhui Huang, Marcel Jahn, Kshitij Kulkarni, Elek Lane, Taylor Madigan, Milan Mossé, Lars Neth, Luca Passi, Richard Pettigrew, Rebecca Rowson, Ezra Rubenstein, Edward Schwartz, David Thorstad, Sarah Vernallis, Yong Xin Hui and Snow Xueyin Zhang. Thanks to anonymous referees for the *Australasian Journal of Philosophy* and *British Journal for the Philosophy of Science* for helpful feedback on the second and third chapter. Thanks to audiences at Berkeley's Richard Wollheim Society, Work in Progress Lunch, dissertation seminar and Formal Epistemology Reading Course (FERC). Also thanks to audiences at Munich Center for Mathematical Philosophy, Universität Bochum, Formal Ethics 2022, Central APA 2022, Berkeley-London Conference 2022, Formal Epistemology Work-

shop 2022, Berkeley-London Conference 2023, 7th Annual CHAI Workshop 2023, Snow Zhang's graduate seminar on Bayesian Epistemology in Fall 2023 and PPE Society Seventh Annual Meeting 2023.

I'm grateful for financial support from a Josephine De Karman Fellowship, a Global Priorities Fellowship from the Forethought Foundation and a Summer Dissertation Writing Grant for Advanced Arts and Humanities and Humanistic Social Sciences Students from Berkeley's Graduate Division. I'm also grateful for the Early Career Conference Programme at Oxford's Global Priorities Institute in summer 2019 which provided a great environment to think about decision theory.

Thanks to Timothy Clarke and R. Jay Wallace for helpful advice on the academic job market and beyond. Thanks to Stefan Gosepath from the Freie Universität Berlin for helpful advice when I first considered studying philosophy in the US.

On a more personal level, thanks to my friends in Germany. Moving to the US to pursue philosophy was (and is) a big adventure and I'm incredibly lucky to have great friends to return to on the other side of the world. Special thanks to Jonathan Scharf and Jonathan Schmitz. Thanks to Basic Channel for the music.

I'm happy to be able to finish this dissertation. Thanks to the urologists at the UKT in Tübingen who saved my life when I was very sick. Thanks to Katharinenhöhe for helping me to become well again.

My biggest thanks go to my family and my partner Morgan. Thank you for always loving and supporting me. Thanks to Conny and Jürgen for supporting me in my strange plan to do philosophy for a living. Thanks to Lars for discussing everything, always. Thanks to Lea for being the best sister in the world. Thanks to Morgan for your love, our years of adventures together since we met in Barry's Theory of Knowledge class, and uncountable hours of talking about everything.

Versions of the second and third chapter are forthcoming:

Sven Neth (forthcoming). "Rational Aversion to Information". In: *British Journal for the Philosophy of Science*. DOI: 10.1086/727772

Sven Neth (forthcoming). "Better Foundations For Subjective Probability". In: *Australasian Journal of Philosophy*

Contents

Abstract **1**

Acknowledgements **iii**

Table of Contents **iv**

1 Introduction **1**
- 1.1 The Bayesian Ideal 2
- 1.2 Non-Ideal Decision Theory 3
- 1.3 Overview 5
- 1.4 What This Dissertation Is 7
- 1.5 What This Dissertation Is Not 8

2 Better Foundations for Subjective Probability **11**
- 2.1 Set-up 13
- 2.2 Ramsey's Method 13
- 2.3 Problems for Ramsey 14
- 2.4 Better Foundations 19
- 2.5 Problems Solved 29
- 2.6 Interpretation 32
- 2.7 Conclusion 35

3 Rational Aversion to Information **37**
- 3.1 Good's Argument 38
- 3.2 Against Good's Argument 44
- 3.3 Value of Information Generalized 57
- 3.4 Conclusion 62

4 Against Coherence **63**
4.1 Lewis' Diachronic Dutch Book 64
4.2 A New Diachronic Dutch Book 67
4.3 The Insurance Analogy 71
4.4 How to (Sometimes) Avoid Sure Loss 75
4.5 Generalizing the Conflict 77
4.6 Against Reflection 82
4.7 Conclusion 86

5 Off-Switching Not Guaranteed **87**
5.1 The Off-Switch Game 88
5.2 The Value of Information 92
5.3 Rational Information Aversion 93
5.4 A Dilemma for Provably Beneficial AI 97
5.5 Conclusion 98

6 Conclusion: Bayesian Modesty **99**

Appendices **100**

A Better Foundations for Subjective Probability **101**

B Rational Aversion to Information **105**

C Against Coherence **110**

Bibliography **128**

List of Figures

1.1 A simple decision problem. 5

3.1 If you care about winning, listen to the weather report. 41
3.2 Ann's decision problem. 51
3.3 Ann's other decision problem. 54

4.1 Diachronic dutch book against Aggu. 66
4.2 Diachronic dutch book against Beatrice. 69
4.3 Fire insurance. 72
4.4 Beatrice without insurance. 73
4.5 Insurance against not conditionalizing. 74
4.6 Diachronic dutch book against uncertain agent. 81

5.1 The Off-Switch Game. 89
5.2 Robbie's decision problem. 90
5.3 Robbie's decision problem with uncertain updating. 95

Chapter 1

Introduction

Alice has incoherent preferences. For example, perhaps her preferences over complicated options are cyclic: she strictly prefers A to B and B to C but also strictly prefers C to A. In decision theory, we often find out what an agents' credences are by looking at their preferences over options. However, when doing so, we assume that agents have perfectly coherent preferences, which rules out agents like Alice. So according to standard decision theory, we cannot ascribe any credences to Alice. But this seems wrong. It would be better if we could make sense of ascribing credences to agents like Alice.

Ann is uncertain about how she will update on some evidence. For example, perhaps she assigns some probability to committing the gambler's fallacy in the future, which means that observing a fair coin landing heads increases her credence that the next coin flip will land tails. Ann is facing a decision and is considering whether she should make her choice now or wait for more information. What should she do? In decision theory, we often assumed that agents are certain about updating. So standard decision theory cannot give any advice to Ann. This seems like a severe restriction of the scope of decision-theoretic advice. It would be better if we could make sense of giving advice to agents like Ann.

The project of my dissertation is to extend the scope of decision theory to 'non-ideal' agents like Alice and Ann. Before explaining the details, it is worth briefly sketching the way decision theory is usually understood.

1.1 The Bayesian Ideal

In decision theory and formal epistemology, we often ask what *ideally rational agents* are like. Standard decision theory (or 'Bayesian decision theory') answers this question as follows:

1. ideally rational agents have probabilistically coherent credences,
2. they always prefer actions which maximize expected utility,
3. they update on any new information by conditionalization.

To explain these assumptions, let us assume, for the purpose of exposition, a finite set of states Ω, which describe all aspects of the world our agent is uncertain about besides their own preferences (Savage 1972). The first assumption demands that our agent's degrees of belief or credences can be modeled by a probability function, which is a function $p : \mathcal{P}(\Omega) \to \mathbb{R}$ satisfying the axioms of probability.[1]

To understand the second assumption, let us assume that there is a set of outcomes $\mathcal{O}$ which contain the ultimate bearers of value for our agent. Let us also assume that our agent's values can be modeled by a utility function $u : \mathcal{O} \to \mathbb{R}$. Then, we can model actions (or 'acts') as functions $f : \Omega \to \mathcal{O}$. The idea is that an action is the kind of thing that has different consequences depending on the (unknown) state of the world. For example, the action of taking an umbrella has the consequence 'stay dry' if it is raining and 'carry an unnecessary umbrella' if it is not raining. The *expected utility* of action f is:[2]

$$\mathbb{E}(f) = \sum_{\omega \in \Omega} p(\{\omega\}) u(f(\omega)).$$

Now the second assumption says that for any actions f and g, a rational agent prefers f to g if and only if f has a higher expected utility than g. While there are subtle questions about the relationship between preference and choice, I will also assume that when there is a finite set of actions, a rational agent will choose one of the actions with maximal expected utility.

[1] The axioms are: (i) non-negativity: $p(X) \geq 0$ for all $X \subseteq \Omega$, (ii) normalization: $p(\Omega) = 1$ and (iii) finite additivity: $p(A \cup B) = p(A) + p(B)$ whenever $A \cap B = \varnothing$. Kolmogorov (1933) suggested the stronger requirement of countable additivity. I briefly discuss countable additivity in Chapter 2 but otherwise stick with finite additivity.

[2] For infinite Ω, we need a slightly more complicated definition.

To understand the third assumption, I will introduce a simple model of learning. I model learning as learning exactly one element of an *evidence partition* $\mathcal{E}$, which is a set of mutually exclusive and collectively exhaustive events with non-zero probability. Imagine, for example, that you are about to take a look at a thermometer. Then, the events in our evidence partition describe different readings of the thermometer. The norm of *conditionalization* says that when an agent starts with credences p and learns event $E \in \mathcal{E}$, they should update their credences from p to $p(\cdot \mid E)$, where $p(A \mid E)$ is the conditional probability of A given E.[3]

The first and second part of the Bayesian picture are often supported by *representation theorems*, which show that any agent whose preferences obey certain axioms of rationality can be represented as expected utility maximizer with probabilistic credences (Ramsey 1926; Savage 1972). The idea that rational agents update on any new information by conditionalization is often supported by *coherence arguments*, which show that agents who plan to update in a different way choose sequences of actions which lead to sure loss over time (Lewis 1999).

Let us suppose for a moment that the Bayesian picture is a correct description of ideal rationality.[4] What are such ideally rational agents like? In many cases, what ideally rational agents prefer and believe depends on the particulars of the situation—their initial credences, their evidence and so on. However, the Bayesian picture entails some structural constraints on preferences and beliefs which hold in general. For example, as Good (1967) shows, ideal Bayesian agents always prefer to learn more information before making a decision given that the information is cost-free (**Value of Learning**). Further, ideal Bayesian agents never choose sequences of actions which lead to sure loss over time (**Coherence**).

1.2 Non-Ideal Decision Theory

What happens if we consider agents who are *not* ideally rational? There is good reason to do so, because it is natural to think of decision theory

[3] I use the ratio definition of conditional probability: $p(A \mid E) = \frac{p(A \cap E)}{p(E)}$ supposing $p(E) > 0$ and set aside other understandings of conditional probability (Hájek 2003).

[4] Talk of 'Bayesian decision theory' or 'Bayesian epistemology' goes back to Bayes (1763), who in this famous posthumously published essay considers the question of how to infer probability from observed frequency (Earman 1992).

not only as a description of ideal agents, but also as a theory that can *give advice* to non-ideal agents. There are, of course, many ways to be non-ideal. I'm interested in a particular kind of non-ideal agent. In Chapter 2, I consider agents who have some irrational preferences. I show how we can extend decision theory to ascribe credences to such agents. In Chapter 3 and Chapter 4, I consider agents who are not certain how they will update on new evidence. I show how we can extend decision theory to give useful advice to such agents.

So the idealizing assumptions I want to relax are:

1. agents have perfectly rational preferences,

2. agents are certain about how they will react to new evidence.

These assumptions are logically independent. You might have some irrational preferences but be certain about how you will react to new evidence. On the other hand, you might have perfectly rational preferences while being uncertain about how you will react to new evidence. And while the first assumption is synchronic—it concerns your preferences right now—the second assumption is diachronic—it concerns decision making over time.

Also note that agents who are not certain about how they will react to new evidence might *in fact* be perfect Bayesian agents. So it's not clear whether agents who are uncertain about how they will update are really non-ideal. My motivation for focusing on such agents is that, as it turns out, the structural constraints which are supposed to follow from the Bayesian picture—**Value of Learning** and **Coherence**—break down when we consider agents who are uncertain about how they will update. This means that good advice to such agents will look very different from the Bayesian ideal.

The picture I have described fits well with a certain naïve way of thinking about decision theory. We face a decision problem: should I take the umbrella or leave it at home? To answer this question, we specify the relevant states of the world—it might rain, it might shine, there could be umbrella muggers—and the subjective probability you assign to these states. We also specify the utility of each pair of state and action. We end up with a decision matrix like the one shown in figure 1.1.

	Take it	Leave it
Rain (.2)	1	-5
Shine (.7)	-1	0
Umbrella muggers (.1)	-10	0

Figure 1.1: A simple decision problem.

The expected utility of taking it is $1 \times .2 - 1 \times .7 - 10 \times .1 = -1.5$ and the expected utility of leaving it is $-5 \times .2 = -1$. As (standard) decision theorists, we can advise you to leave the umbrella at home. When giving this advice, it seems we do not presuppose that you already prefer the action with the higher expected utility. You might initially prefer to take the umbrella. After all, decision making under uncertainty is hard.

1.3 Overview

Here is an overview of the chapters of this dissertation.

1.3.1 Better Foundations For Subjective Probability

To give decision-theoretic advice to an agent, we first need to know what their subjective probabilities are. How do we ascribe subjective probabilities? The standard answer, pioneered by Ramsey (1926) and Savage (1972), is to use *representation theorems*, which show that an agent whose preferences obey certain axioms can be represented as an expected utility maximizer relative to a unique probability function.

However, standard representation theorems presuppose that the agent under consideration has perfectly rational preferences. This means that these representation theorems cannot underwrite the naïve picture of decision theory sketched above, where we draw on the agent's probabilities to give advice which potentially corrects our agent's preferences.

I show that we can do better. I present a representation theorem building on Savage (1972) and Krantz et al. (1971) which can be used to measure or define an agent's subjective probabilities given weak rationality axioms which allow irrational preferences. The key idea is to start with comparative probability judgments and to construct a unique probability function which represents these comparative judgments. This representation theorem makes room for useful decision-theoretic advice.

1.3.2 Rational Aversion to Information

After making sure we can give decision-theoretic advice to non-ideal agents, I turn to the question *what kind of* advice we should give. There are many ways to be non-ideal and I focus on one kind of deviation from the classical Bayesian ideal: agents who are *modest*, which means they are uncertain about how they will update on new information. Modest agents might, as a matter of fact, never deviate from conditionalization, but they are not sure of this.

It is reasonable to be modest. Modesty follows from plausible principles of Bayesian epistemology, which urge us to assign non-zero probability to all empirical propositions and proportion our belief to the available evidence. So Bayesian epistemology suggests that we should not assume we are perfect Bayesian agents. Whether we are perfect Bayesian agents is itself an empirical question and it is reasonable to assign some credence to the possibility that we are not.

For modest agents, maximizing expected utility sometimes requires rejecting learning free and relevant information before making a decision. Good's theorem, which is often glossed by saying that 'expected utility maximizers are never worse off by learning more information', does not apply to modest agents. This is because Good assumes not only that the agent under consideration maximizes expected utility, but also that they are certain of updating by conditionalization.

1.3.3 Against Coherence

Another structural constraint implied by the Bayesian ideal is **Coherence**, which says that an agent will never make a sequence of choices which, over time, yields sure loss. I show that for modest agents, maximizing expected utility sometimes leads to incoherence. While some might take this as a reason against modesty, I argue that we should embrace incoherence instead, drawing on an analogy with purchasing insurance. The upshot is that like **Value of Learning**, **Coherence** is *normatively fragile* and does not constitute robust advice for non-ideal agents. We should be suspicious of arguments which purport to derive other rational requirements from **Coherence**.

The overall lesson is that **Value of Learning** and **Coherence** follow from expected utility maximization only if we assume that agents are certain of updating by conditionalization. Since this requirement will fail for non-ideal agents and even for agents which are in fact perfect Bayesians but

uncertain about this, **Value of Learning** and **Coherence** are not robust normative requirements.

1.3.4 Off-Switching Not Guaranteed

One application of decision theory is to inform the design of AI systems, and an important problem in the design of AI systems is how we can make sure that AI systems will always defer to us if we want to switch them off (Russell 2019). Hadfield-Menell et al. (2017) propose the Off-Switch Game, a simple model of Human-AI cooperation in which AI agents always defer to humans because they are uncertain about our preferences. I show how deference can fail if the AI agent is not certain of updating by conditionalization. So my framework has important consequences for applied decision theory.

1.4 What This Dissertation Is

Here are two ways to think about the overall picture which emerges from this dissertation.

We can distinguish between normative and interpretative decision theory (Buchak 2017, p. 789). In normative decision theory, we take agent's credences and utilities for granted and ask what they should do given that they have these credences and utilities. As noted above, the standard advice is that a rational agent should maximize expected utility relative to their credence function and utility function. In interpretative decision theory, we do not take credences and utilities for granted. Rather, we start with an agent's preferences over actions and use the assumption that the agent obeys some decision-theoretic principle to 'reverse engineer' their credences and utilities from their preferences. The standard story of interpretative decision theory, pioneered by Ramsey (1926), is that we can use the assumption that an agent maximizes expected utility to infer their credences and utilities from their preferences. We can think of this as a project of 'radical interpretation': figure out what an agent believes and desires from scratch, merely by looking at their preferences.

We can think of Chapter 2 as extending interpretative decision theory to non-ideal agents. As I will show, we can use assumptions much weaker than expected utility maximization to infer credences from preferences. In contrast, Chapter 3 and Chapter 4 extend normative decision theory to non-

ideal agents. In these chapters, I take credences and utilities for granted and give advice what agents should do if they are uncertain about their updating behavior. So one way to think about the overall upshot of this dissertation is as showing how to do both interpretative and normative decision theory for non-ideal agents.

Here is another way of thinking about it. In this dissertation, I discuss three very influential ideas in decision theory. First, representation theorems, which show how to define or infer subjective probability by looking at preferences (Ramsey 1926; Savage 1972). Second, value of information, the idea of measuring the value of learning by its expected impact on future decisions (Good 1967). Third, coherence, the idea that rational agents should not make sequences of choices which lead to sure loss (de Finetti 1937; Lewis 1999).[5] In all three cases, considering non-ideal agents leads to interesting results, although these results pull us in somewhat different directions. We can generalize representation theorems to non-ideal agents, so representation theorems can serve as a solid foundation for decision theory even if we relax idealizing assumptions. In contrast, allowing agents to be uncertain about updating undermines the idea that the expected value of information is non-negative and that we should be coherent.

Here is a worry for the project of non-ideal decision theory. There are many ways to relax idealizations of decision theory, which can make it seem arbitrary to focus on one of them rather than another. The big-picture upshots discussed above help to respond to this worry. My way of doing non-ideal decision theory is fruitful because it engages with ideas which form the core of decision theory: interpretative and normative decision theory, representation theorems, value of information and coherence.

1.5 What This Dissertation Is Not

As noted above, there are many ways to be non-ideal. Thus, there are many alternative ways to do non-ideal decision theory which I will not discuss here.

Most importantly, for the purpose of this dissertation, I will assume that

[5] Arguably, all three ideas are due to Ramsey. In addition to introducing representation theorems, Ramsey (1926) suggests that we can use coherence as a foundation for probabilism (while not explicitly drawing the connection between coherence and conditionalization). In a posthumously published note, Ramsey (1990) sketches a version of Good's theorem.

standard decision theory correctly describes ideally rational agents. In particular, I will assume that rationality requires precise credences and maximizing expected utility. There are alternative conceptions of ideal rationality, for example alternative decision theories (Buchak 2013) or imprecise credences (Joyce 2010). Imprecise credences are sometimes motivated by thought that standard decision theory is too demanding in requiring us to assign precise real-valued probabilities. But as I will show in Chapter 2, we can provide foundations for precise subjective probabilities which apply to non-ideal agents. I am sympathetic to the idea that even ideally rational agents are not required to maximize expected utility. My arguments in Chapter 3 and Chapter 4, which show that rational agents can be required to reject free information and suffer sure loss, could be recast with the weaker assumption that expected utility maximization is rationally permissible. But to keep it simple, I will assume that rationality requires expected utility maximization.

I will not consider computationally limited agents, the study of which is often called 'bounded rationality'. There is a rich research tradition exploring these ideas in psychology, economics and artificial intelligence (Simon 1976; Cherniak 1986; Russell and Subramanian 1995; Gigerenzer and Goldstein 1996; Aumann 1997; Griffiths, Lieder, and Goodman 2015; Icard 2018; Thorstad 2022b; Thorstad forthcoming). One of the core ideas of this tradition is to move from rational requirements on preferences to rational requirements on processes of inquiry and decision-making.[6] While I think this is a fruitful approach, we will see that considerations of non-ideal agents lead to rich and interesting consequences even if we stick with the formal framework of standard decision theory and relax rationality assumptions within this framework.

Relatedly, I will not talk about the problem of logical omniscience (Savage 1967; Stalnaker 1991). My framework has no room to model agents who are uncertain about logical truths. I will also set aside issues of unawareness (Steele and Stefánsson 2021). Again, while I think these problems are pressing, non-ideal decision theory is interesting even if we assume logical omniscience and focus instead on agent's empirical uncertainty about their own future updating behavior. One advantage of going this route is that it is hard to come up with a simple framework to model agents who are uncertain

[6]While not working within the bounded rationality tradition, Harman (1986) proposes an account in a similar spirit, focusing on the process of reasoning instead of rational requirements on beliefs and preferences.

about logical truths.[7]

Some philosophers have framed theories of non-ideal agents in terms of the question whether and why these agents should *approximate* normative ideals (Bona and Staffel 2018; Staffel 2019). While this is a fruitful question, agents do not always have the option to approximate the ideal. I focus on what agents should do if they don't have the option to become 'better Bayesians'.

I will also set aside issues of uncertain evidence and assume that when you learn something, this can be modeled as becoming certain of some proposition. According to Jeffrey (1957), experience can shift our probabilities even though there is no proposition we become certain of, which leads to a rich framework of updating rules for such learning experiences (Diaconis and Zabell 1982). As Huttegger (2017) argues, Jeffrey's move can be understood as a way of de-idealizing the standard Bayesian model of learning. However, I set such complications aside and focus on relaxing another assumption of the standard model: you are certain how learning a proposition with certainty will impact the rest of your credence function.

Furthermore, I will set aside questions about the relationship between credence and 'full belief' (Buchak 2014; Leitgeb 2014; Jackson 2020; MacFarlane forthcoming). Such questions are potentially relevant to non-ideal theorizing since one might think that we ascribe and reason with beliefs because they are more tractable than credences. However, even if we confine our attention to credences, there are more than enough questions to ponder.

I will set aside the dispute between causal and evidential decision theory (Nozick 1969; Lewis 1981; Joyce 1999). Throughout this dissertation, I assume that actions are causally and probabilistically independent of states, so that causal and evidential decision theory will agree in their verdicts. Finally, I will set aside problems of self-locating uncertainty (Lewis 1979).

With these preliminaries out of the way, let us now begin.

[7]Hacking (1967), Garrabrant et al. (2016), Elga and Rayo (2022), and Skipper and Bjerring (2022) propose different ways to approach the problem of logical omniscience in a Bayesian framework.

Chapter 2

Better Foundations for Subjective Probability

In philosophy, psychology and economics, we often ascribe subjective probability or credence to people. For example, we might say that Alice's subjective probability that it will rain tomorrow is .3. What is the basis for such ascriptions of subjective probability?[1]

If we want to find out Alice's subjective probability that it will rain tomorrow, one natural idea is to look at which bets Alice is willing to accept. If I offer Alice a bet which pays one dollar if it rains tomorrow, how much is this bet worth to her? There is a long tradition in decision theory inspired by this idea, going back to Ramsey (1926). In this tradition, we assume that an agent satisfies certain principles of rationality and then *define* or *measure* subjective probability in terms of preferences.[2]

But what if Alice does not satisfy the rationality assumptions required by Ramsey? Ramsey assumes that the agent under consideration is an expected utility maximizer and there are good reasons to doubt that real-life agents maximize expected utility. Now perhaps this means that real-life agents do not have any subjective probabilities or that we cannot measure them. However, there is an alternative option: we can provide decision-theoretic foundations for subjective probability with weaker rationality assumptions.

[1] A version of this chapter is forthcoming in the *Australasian Journal of Philosophy*.

[2] Buchak (2017) calls this 'interpretative decision theory': we use decision-theoretic principles to interpret an agent's mental states in a process of 'radical interpretation' (Davidson 1973; Lewis 1974). As I briefly discuss in appendix A, the idea of betting as a guide to degrees of belief can be traced back at least as far as Kant.

I introduce a representation theorem building on Savage (1972) and Krantz et al. (1971) which connects subjective probability to preference with much weaker rationality assumptions than standard representation theorems. In particular, I allow agents to not maximize expected utility, to violate stochastic dominance and to consider most options incomparable. The key idea is to start with comparative probability judgments and to construct a unique probability function which represents these comparative judgments. My representation theorem has important philosophical upshots: it makes sense of how we can ascribe precise subjective probability to partly irrational agents and how decision theory can provide useful advice.

As suggested above, there are two ways of understanding the project of grounding ascriptions of subjective probability in preferences. First, one might attempt to define subjective probability in terms of preference. On this view, to say that Alice's subjective probability that it will rain tomorrow is .3 just means that Alice is willing to accept certain bets. So ascriptions of subjective probability are a 'representational device' to talk about something more fundamental: Alice's preferences. Let us call this view *constructivism*. Second, one might think that while subjective probability is not reducible to preference, we can measure subjective probability by observable preferences. Call this view *realism*. I will mostly remain neutral between constructivism and realism. Like Ramsey's approach, my representation theorem can be interpreted as defining or measuring subjective probability in terms of preference. However, I will later suggest that the theorem naturally fits with an intermediate position between constructivism and realism: comparative probability is psychologically real but numerical probability functions are merely 'representational devices' for talking about comparative probability.

Here is the plan. I start by introducing some terminology (2.1) and explain Ramsey's method for measuring subjective probability (2.2). I discuss some problems for Ramsey's method (2.3). I introduce better foundations for subjective probability (2.4) and explain how they overcome the problems for Ramsey's method (2.5). I finish by sketching two ideas suggested by the representation theorem (2.6): the view that comparative probability is more fundamental than numerical probability and a subjectivist version of the classical interpretation of probability.

2.1 Set-up

We have a set Ω of *states* which describe the world apart from our agent's preferences and a σ-algebra $\mathscr{F}$ of subsets of Ω which are called *events*.[3] For any $X \in \mathscr{F}$, we denote the relative complement of X in Ω by X^C. We have a set $\mathscr{O}$ of *outcomes* which contain everything our agent cares about.

Following Savage (1972), *acts* are functions from states to outcomes. An act is *finite-valued* iff it only takes finitely many different outcomes. Our *act space* $\mathscr{A}$ is the set of all finite-valued acts. This means that we can write each act $f \in \mathscr{A}$ as $\{o_1, E_1; ...; o_n, E_n\}$, where events $E_1, ..., E_n$ are pairwise disjoint sets and their union is Ω and for each E_i with $1 \leq i \leq n$, o_i is the unique outcome $o_i \in \mathscr{O}$ such that $f(\omega) = o_i$ for all $\omega \in E_i$.

I write '$\succsim$' for our agent's preference ordering over acts, a binary relation on $\mathscr{A}$. The intended interpretation of $f \succsim g$ is that our agent weakly prefers f to g. Strict preference ($\succ$) and indifference ($\sim$) are defined in the usual way.[4] For each outcome $o \in \mathscr{O}$, the *constant act yielding* o, written $\underline{o}$, is the act which assigns o to all $\omega \in \Omega$. I define, for any $o, o' \in \mathscr{O}$, $o \succsim o'$ iff $\underline{o} \succsim \underline{o'}$. I use the term 'option' to talk about both acts and outcomes.

2.2 Ramsey's Method

Ramsey (1926) proposes axioms on preferences which imply that our agent is representable as expected utility maximizer.[5] This means that there is some probability function and utility function such that our agent always prefers acts with higher expected utility.[6] Ramsey then gives us a way to construct or infer our agent's utility function without already knowing our agent's probability function.

Once we have the utility function, Ramsey pins down the subjective probability of any event $E \in \mathscr{F}$ as follows. First, we find three outcomes $b, m, w \in \mathscr{O}$ (best, medium and worst) such that our agent strictly prefers

[3] A *σ-algebra* on Ω is a set of subsets of Ω which contains Ω and is closed under complementation and countable unions.

[4] So $f \succ g \iff (f \succsim g) \wedge \neg(g \succsim f)$ and $f \sim g \iff (f \succsim g) \wedge (g \succsim f)$.

[5] Jeffrey (1990, Ch. 3) and Bradley (2004) reconstruct Ramsey's reasoning. Ramsey originally used a different framework which does not distinguish states and outcomes, while I am reconstructing Ramsey's reasoning in terms of the Savage framework.

[6] Relative to probability function $p : \mathscr{F} \to [0, 1]$ and utility function $u : \mathscr{O} \to \mathbb{R}$, the *expected utility* of act $f = \{o_1, E_1; ....; o_n, E_n\}$ is $\mathbb{E}_{u,p}(f) = \sum_{i=1}^{n} p(E_i)u(o_i)$.

b over w and is indifferent between getting m for certain and an act which yields b if E happens and w otherwise:

$$b \succ w, \tag{2.1}$$

$$m \sim \{b, E; w, E^C\}. \tag{2.2}$$

Then we use the assumption that our agent maximizes expected utility to infer that $p(E) = \frac{u(m)-u(w)}{u(b)-u(w)}$.[7] So for Ramsey, subjective probabilities are betting odds. Since event E was arbitrary, we can use Ramsey's method to uniquely pin down the subjective probability of all events. Depending on whether we accept constructivism or realism, we can think of this as a definition of subjective probability in terms of preferences or as a way to measure subjective probability by preferences.

Ramsey had a lasting influence on decision theory.[8] Savage (1972) also lays down axioms on the preference relation and proves a representation theorem which shows that any agent obeying these axioms can be represented as expected utility maximizer with a unique probability function. Many of the problems for Ramsey's approach I will discuss below generalize to Savage's representation theorem. However, as we will see later, the work of Savage holds key insights for an alternative approach to measure subjective probability.[9]

2.3 Problems for Ramsey

I now turn to explain why Ramsey's method is not an adequate foundation for subjective probability.

[7] Proof: since our agent maximizes expected utility, (2.1) and (2.2) entail that $u(m) = u(w) + p(E)(u(b) - u(w))$. Therefore, $u(m) - u(w) = p(E)(u(b) - u(w))$, so $p(E) = \frac{u(m)-u(w)}{u(b)-u(w)}$. The utility function is unique up to positive affine transformation so $p(E)$ is unique.

[8] Fishburn (1981) provides a great survey of decision theory after Ramsey. Misak (2020) places Ramsey's work in its broader intellectual context.

[9] Jeffrey (1990) develops a different framework for decision theory in which states, acts and outcomes are all propositions. However, Jeffrey's axioms do not pin down a unique probability function.

2.3.1 Strong Rationality Assumptions

Ramsey assumes that the agent under consideration is an expected utility maximizer. However, there are good reasons to think that real-life agents are not expected utility maximizers. Suppose you only care about money and choose between the following two lotteries:[10]

1. One million dollars for certain.

2. 89 % chance of winning one million dollars, 10 % chance of winning five million dollars, 1 % chance of winning nothing.

You also choose between the following two lotteries:

3. 89 % chance of winning nothing, 11 % chance of winning one million dollars.

4. 90 % chance of winning nothing, 10 % chance of winning five million dollars.

If you strictly prefer (1) over (2) and (4) over (3), your preferences are incompatible with expected utility maximization (Allais 1953).[11] However, real-life agents sometimes exhibit this pattern of preferences (Oliver 2003). My point here is not that this pattern of preferences is rationally permissible, as argued by Buchak (2013). Rather, my point is that real-life agents apparently have such 'Allais-preferences'. Therefore, we cannot use Ramsey's method to define or measure their subjective probabilities. However, it still seems like such agents have subjective probabilities—after all, they are presumably *using* their subjective probabilities to reason that (1) is better than (2) and (4) is better than (3). So Ramsey's method is not a good foundation for subjective probability. To make this vivid, imagine you find out that Alice has the preferences described above. Should you conclude that Alice cannot have any subjective probabilities or that there is no way for us to find out what these probabilities are? I think not.

[10] A *lottery* is a probability distribution over outcomes and can be realized by multiple acts.

[11] If you strictly prefer (4) over (3), $.1u(\$5\text{ Million}) > .11u(\$1\text{ Million})$. So, adding the same term on both sides, $.1u(\$5\text{ Million}) + .89u(\$1\text{ Million}) > .11u(\$1\text{ Million}) + .89u(\$1\text{ Million})$, which means that $.1u(\$5\text{ Million}) + .89u(\$1\text{ Million}) > u(\$1\text{ Million})$. But if you strictly prefer (1) over (2), $u(\$1\text{ Million}) > .89u(\$1\text{ Million}) + .1u(\$5\text{ Million})$.

People also sometimes choose *stochastically dominated* options. Option A stochastically dominates option B if for every outcome $o \in \mathscr{O}$, the probability that A yields an outcome weakly preferred to o is greater than or equal to the probability that B yields an outcome weakly preferred to o. It is generally agreed that you should:

> **Respect Stochastic Dominance.** If f stochastically dominates g, then $f \succsim g$.

This principle follows from many normative decision theories, such as expected utility theory, risk-weighted expected utility and others.[12] However, empirical studies show that people sometimes violate this principle. Consider the following two lotteries:

5. 5% chance of \$12, 5% chance of \$14, 90% chance of \$96.

6. 10% chance of \$12, 5% chance of \$90, 85% chance of \$96.

It is not hard to see that (5) stochastically dominates (6).[13] Nonetheless, in a study reported by Birnbaum and Navarrete (1998), most subjects chose (6) over (5). This is presumably because they rely on quick but imperfect heuristics in their decision making. Again, my claim is not that these preferences are rational but only that real-life agents have such preferences. Therefore, we cannot use Ramsey's method to define or measure their subjective probabilities. But again, while it might be irrational to have preferences which violate stochastic dominance, such preferences do not seem to preclude agents from having subjective probabilities.

Finally, people sometimes regard options as *incomparable* in value. Consider, for example, the choice between a career as a doctor and a career as a rock star. Both career choices can lead to a fulfilling and valuable life. However, what makes them valuable is radically different. It is difficult to see how one could compare the two options. Someone could reasonably think that one is not better than the other but neither are they exactly equally good (Chang 2002).

[12] Buchak (2013) and Tarsney (2020) defend stochastic dominance. Bader (2018) points out the wide applicability of stochastic dominance reasoning even if outcomes are incomparable. Russell (forthcoming) discusses some problems arising in such a setting.

[13] Both lotteries are sure to pay at least \$12. The probability of winning at least \$14 is 95% in (5) and 90% in (6). The probability of winning at least \$90 is 90% in (5) and (6).

Incomparability arises both at the level of outcomes and acts. It is natural to understand the career choice example in terms of incomparable outcomes. In contrast, a second kind of incomparability might arise because it is too difficult to compare acts even if all of their outcomes are comparable. Suppose, for example, that you like more money rather than less. When faced with a choice between two complicated investment portfolios, you might nonetheless not have any preference between them because it is too difficult for you to reason about the decision problem. Again, my claim is not that incomparability is rational, but merely that real-life agents sometimes have such preferences.[14]

Expected utility theory has no room for incomparability. This is because your utility function assigns a real number to each outcome and so renders all outcomes comparable. Each act is ranked by its expected utility so all acts are comparable as well. Since real-life agents sometimes regard both outcomes and acts as incomparable, their preferences cannot be represented as expected utility maximization. However, it is not plausible that incomparability precludes agents from having subjective probabilities.

Here is the upshot. There are good reasons to think that real-life agents are not expected utility maximizers. Therefore, we cannot use Ramsey's method to define or measure their subjective probabilities. Some might take this as reason to embrace a kind of nihilism: such agents do not have subjective probabilities or there is no way to measure what they are. A better response is to provide foundations for subjective probability which apply even to agents which fail to maximize expected utility, do not respect stochastic dominance and consider some options incomparable. One might still think that we should model the beliefs of real-life agents by something other than precise probability functions. However, we can make room for irrational preferences without giving up decision-theoretic foundations for precise subjective probability.

[14]Many have defended the stronger claim that incomparability can be rational. Joyce (1999, p. 102) writes that "a decision maker can be perfectly rational even when her preferences do not satisfy the completeness axiom". Aumann (1962, p. 446) writes that "[o]f all the axioms of utility theory, the completeness axiom is perhaps the most questionable". Similar points are made by Hare (2010), Bales, Cohen, and Handfield (2014), Schoenfield (2014), Bader (2018), and Sen (2018).

2.3.2 No Useful Advice

The standard decision-theoretic advice is to maximize expected utility relative to your subjective probability function and utility function. For this advice to be useful, we first need to figure out what your probability function *is*.[15] However, if we define or measure your probability function on the assumption that you maximize expected utility, the advice to maximize expected utility can never be useful. Therefore, decision theory cannot play the role of giving useful advice.[16]

This puzzle arises on both constructivism and realism. For constructivists, your probability function is defined in terms of preferences which satisfy certain axioms. If you violate the axioms, you simply do not *have* a probability function and the advice to maximize expected utility is meaningless. For realists, you might still have a probability function if you violate the axioms. However, standard representation theorems give us no way to infer what this probability function is, so we cannot use decision theory to give useful advice.[17]

One reaction to this problem is to say that the only advice decision theory provides is: 'obey the axioms!'. On this view, decision theory is merely a theory of consistency (Dreier 1996; Okasha 2016). While I have no knockdown objection to this position, it is unattractive because it makes decision theory largely irrelevant to non-ideal agents like us who are pretty much guaranteed to violate some normative principle of decision making. It would be better if we could make sense of how decision theory can provide useful advice to partly irrational agents. As I will show, we can indeed make sense of this, which considerably weakens the plausibility of this response.

[15]The same point applies to the utility function but I focus on subjective probability.

[16]Resnik (1987, p. 99) writes, about representation theorems in decision theory: "the theorem can be applied only to those agents with a sufficiently rich preference structure; and if they have such a structure, they will not need utility theory—because they will already prefer what it would advise them to prefer". Meacham and Weisberg (2011) and Easwaran (2014) deploy similar arguments. Beck and Jahn (2021) also discuss the question of how decision-theoretic models can provide useful advice.

[17]This puzzle also arises for non-standard decision theories such as risk-weighted expected utility theory (Buchak 2013) and weighted-linear utility theory (Bottomley and Williamson forthcoming). On these theories, we also need an independent grasp on your subjective probability function, your utility function and possibly other functions like your risk function in order for the theory to provide useful advice.

2.3.3 Dependence on Utility

Ramsey defines subjective probability as ratio of utilities. This requires a very rich space of outcomes. For Ramsey, outcomes must allow for continuous gradations of value.[18] However, it seems like agents can have subjective probabilities while not making such fine-grained distinctions of value. We could even imagine agents who do not have a utility function at all but merely an ordinal ranking of outcomes. For example, we can imagine an agent which only distinguishes between two outcomes, GOOD and BAD. Ramsey must deny that such an agent could have subjective probabilities or that we can find out what they are. This seems implausible.

More broadly, Ramsey's approach gives utility a certain kind of priority over subjective probability. But you might think that subjective probability is conceptually independent of utility. It would be great to disentangle the assumptions needed to measure subjective probability from strong assumptions about the structure of value, for example that the value of all outcomes is comparable and that value can be measured by a real-valued utility function. Such assumptions about value have seemed implausible to many philosophers and it would be great to have foundations for subjective probability which do not rely on them.

2.4 Better Foundations

We can provide better foundations for subjective probability. I introduce and explain a representation theorem building on Savage (1972) and Krantz et al. (1971) which yields a unique probability function representing our agent's beliefs on weak rationality assumptions. The key idea is to start with *comparative probability judgments* and to construct a unique probability function which represents these comparative judgments.

2.4.1 Comparative Probability

What does it mean to think that one event is more probable than another? Savage (1972) proposes to define comparative probability judgments in terms of certain kinds of preferences. Suppose our agent strictly prefers outcome b

[18]For example, Fishburn (1981, p. 152) writes that in Ramsey's approach, the set of outcomes "must be infinite and give arbitrarily fine gradations in utility".

over outcome w. Now the intuition is that our agent *prefers the better prize on the more probable event*. So if our agent prefers the act $\{b, X; w, X^C\}$ over $\{b, Y; w, Y^C\}$, this means that our agent believes that event X is at least as likely as event Y, written $X \succcurlyeq Y$. So we can use acts of the form $\{b, X; w, X^C\}$, where $b \succ w$, to define or infer our agent's comparative probability judgments. Let us call these *test acts*.

We define the following relation $\succcurlyeq$ on $\mathscr{F}$:

Definition 1. *$X \succcurlyeq Y$ iff $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ for some $b, w \in \mathscr{O}$ with $b \succ w$.*

Strict comparative probability ($\succ$) and indifference ($\approx$) are defined in the standard way.[19]

We can understand Savage's proposal in two ways. For constructivists, comparative probability reduces to preferences. (This is Savage's own view.) For realists, comparative probability does not reduce to preferences, but we can use preferences to measure comparative probability judgments. The core proposal of this paper is compatible with both ways of understanding comparative probability. However, I think that realism about comparative probability is ultimately more plausible and can tell a better story about some of the axioms below. I will return to this issue later.

2.4.2 Axioms

The first axiom ensures that the comparative probability ordering does not depend on our particular choice of outcomes:

> **Outcome Independence**. For all $X, Y \in \mathscr{F}$, if $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ for some $b, w \in \mathscr{O}$ such that $b \succ w$, then $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ for all $b, w \in \mathscr{O}$ such that $b \succ w$.

You would violate this axiom if you prefer to bet one dollar on event X rather than event Y but you also prefer to bet two dollars on Y rather than X. In this case, we cannot elicit stable comparative probability judgments from your preferences.

[19] $X \succ Y \iff (X \succcurlyeq Y) \wedge \neg(Y \succcurlyeq X)$ and $X \approx Y \iff (X \succcurlyeq Y) \wedge (Y \succcurlyeq X)$. In a slight abuse of notation, I use the same symbol ($\succ$) for both strict preference and strict comparative probability. This way to link comparative probability to preferences is standard (Fishburn 1986; Icard 2016).

The next axiom demands that our agent is not indifferent among all outcomes:

Non-Degeneracy. There are outcomes $b, w \in \mathscr{O}$ with $b \succ w$.

What is the status of this axiom? Does the existence of subjective probability really require that you are not indifferent between all outcomes? Eriksson and Hájek (2007) point out that we can imagine a Zen monk who is indifferent between all outcomes but nonetheless has subjective probabilities. Thus, there are problems with **Non-Degeneracy** understood along constructivist lines. However, if we are realists, we can accept **Non-Degeneracy** as a condition under which we can measure comparative probability. The Zen monk might have subjective probabilities, but if they are really indifferent among everything, there is simply no way for us to find out what these subjective probabilities are. So we can think of this axiom as a *structure axiom* which ensures that preferences are rich enough to measure subjective probability.[20]

Here is the third axiom:

> **Restricted Ordering**. The relation $\succsim$ restricted to test acts with the same outcomes is complete and transitive. This means for any $b, w \in \mathscr{O}$ with $b \succ w$, for all $X, Y \in \mathscr{F}$ we have either $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ or $\{b, Y; w, Y^C\} \succsim \{b, X; w, X^C\}$.[21] And if $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ and $\{b, Y; w, Y^C\} \succsim \{b, Z; w, Z^C\}$, then $\{b, X; w, X^C\} \succsim \{b, Z; w, Z^C\}$.

I do not constrain the preference relation in general to be complete and transitive, which leaves room for incomparability.

Why accept **Restricted Ordering**? Given our definition of comparative probability, **Restricted Ordering** requires that the comparative probability judgments of our agent are complete and transitive. There are reasons to be skeptical of both.[22] For proponents of imprecise credences, rejecting completeness is particularly natural. Perhaps you have some opinion about

[20] Joyce (1999, p. 82) distinguishes structure axioms and rationality axioms.

[21] Two test acts with different outcomes needn't be comparable if the outcomes themselves are incomparable. Thanks to an anonymous referee for bringing this issue to my attention.

[22] Fishburn (1986, p. 339) discusses examples in which comparative probability judgments violate completeness and transitivity. Ding, Holliday, and Icard (2021) study logics for comparative probability without completeness.

how likely it is that there is life on Mars and some opinion about how likely it is to rain tomorrow but no opinion about which is more likely. This seems particularly plausible if we consider agents which are not perfectly rational.

In response, remember that I want to explain how it is possible to ascribe *precise* credences to agents with some irrational preferences. For this reason, I will not consider agents whose comparative probability judgments fail to be complete and transitive. Such agents fall outside of the scope of my project.

The next two axioms are where the main action is. Let us begin with:

> **Certain Prize**. For any $b, w \in \mathscr{O}$, if $b \succ w$, then for any $X \in \mathscr{F}$, $\underline{b} \succsim \{b, X; w, X^C\}$ and $\{b, X; w, X^C\} \succsim \underline{w}$.

This principle states a plausible minimal rationality condition. It says that if you strictly prefer b to w, then you must weakly prefer getting b for certain to an act which yields b if X happens and w otherwise. Further, you must weakly prefer this act to getting w for certain.

While **Certain Prize** is quite weak, it is possible to imagine agents which violate this axiom. For example, agents might prefer a risky option over a sure thing because they enjoy the thrill of gambling.[23] Relatedly, **Certain Prize** might be violated by agents who prefer randomization (Icard 2021). One response to this concern is to make more fine-grained distinctions among outcomes (Dreier 1996). For example, a prize obtained for sure would be a different outcome from the same prize obtained by a risky gamble. However, this move threatens to trivialize decision-theoretic norms. So it is best to concede that while the axiom is weaker than rationality axioms in standard representation theorems, it still makes substantive demands which some agents might violate.

Do agents which violate **Certain Prize** not have subjective probabilities? This is not very plausible. After all, it is precisely their subjective probabilities which lead them to prefer the risky option. It is more plausible to think that if agents love the thrill of gambling, it might be difficult to determine their subjective probabilities from their preferences. As I show below, **Certain Prize** is a necessary condition for the agent's comparative probability judgments to be representable by a probability function, so measuring the subjective probability of agents which violate **Certain Prize** would require a fundamentally different approach to measuring subjective probability.

Here is another key axiom:

[23] Thanks to an anonymous referee for raising this objection.

> **Alternative Prize**. For any $X, Y, Z \in \mathscr{F}$ and $b, w \in \mathscr{O}$, if $b \succ w$ and Z is such that $X \cap Z = Y \cap Z = \varnothing$, then $\{b, X; w, X^C\} \succ \{b, Y; w, Y^C\}$ iff $\{b, X \cup Z; w, (X \cup Z)^C\} \succ \{b, Y \cup Z; w, (Y \cup Z)^C\}$.

Alternative Prize says the following. Suppose you strictly prefer b over w and you prefer $\{b, X; w, X^C\}$ over $\{b, Y; w, Y^C\}$. Now we modify both acts as follows: You also get b if some event Z disjoint from both X and Y happens. Now you should prefer $\{b, X \cup Z; w, (X \cup Z)^C\}$ to $\{b, Y \cup Z; w, (Y \cup Z)^C\}$. This reasoning also works backwards. **Alternative Prize** has a clear interpretation in terms of probability. If you prefer $\{b, X; w, X^C\}$ to $\{b, Y; w, Y^C\}$, you think that X is at least as likely as Y. Therefore, $X \cup Z$ must be at least as likely as $Y \cup Z$ given that Z is disjoint from both X and Y. So you should prefer $\{b, X \cup Z; w, (X \cup Z)^C\}$ to $\{b, Y \cup Z; w, (Y \cup Z)^C\}$ since you want the better prize on the more probable event.

You might violate **Alternative Prize** if you have credences which are not additive and represented by an alternative formalism like Dempster-Shafer functions or ranking theory.[24] But my goal is to provide foundations for ascribing subjective *probability* to partly irrational agents. So agents modeled by such formalisms fall outside the scope of my project.[25] It would be desirable to have more general foundations for measuring belief which apply to agents with non-probabilistic credences, but I will not consider such agents here.

My axioms on the preference relation are necessary and sufficient for the comparative probability ordering to be a *qualitative probability* (de Finetti 1931):

Definition 2. *A binary relation $\succcurlyeq$ on $\mathscr{F}$ is a* qualitative probability *iff for all $X, Y, Z \in \mathscr{F}$:*

1. *$\succcurlyeq$ is complete and transitive (Ordering),*
2. *$\Omega \succcurlyeq X \succcurlyeq \varnothing$ (Boundedness),*
3. *$\Omega \succ \varnothing$ (Non-Triviality),*

[24] Ellsberg (1961) gives an example of preferences which violate **Alternative Prize**.

[25] Titelbaum (2022, Ch. 14.3) gives a brief introduction to Dempster-Shafer functions and Spohn (2012) discusses ranking theory. As both authors note, it is unclear how these alternatives to probability interact with decision making, which is a reason to set them aside for our purposes. Thanks to an anonymous referee for the suggestion to consider these frameworks.

4. *if* $X \cap Z = Y \cap Z = \varnothing$*, then* $X \succ Y \iff X \cup Z \succ Y \cup Z$ *(Qualitative Additivity).*

Theorem 1. *The preference relation* $\succsim$ *satisfies* ***Outcome Independence***, ***Non-Degeneracy***, ***Restricted Ordering***, ***Certain Prize*** *and* ***Alternative Prize*** *if and only if the comparative probability ordering* $\succcurlyeq$ *is a qualitative probability.*

A proof is provided in appendix A. I follow Savage's definition of comparative probability. Savage also assumes **Outcome Independence** and **Non-Degeneracy**. The key difference is that Savage uses much stronger axioms to derive the result that the comparative probability ordering is a qualitative probability. Instead of **Restricted Ordering**, Savage assumes that preferences are complete and transitive, which rules out incomparable options. This strong assumption is unnecessary to establish that the comparative probability ordering is complete and transitive. It suffices to assume that a small fragment of the preference relation is complete and transitive.

Further, Savage appeals to the 'Sure-Thing Principle' in order to establish that the comparative probability ordering satisfies Boundedness, Non-Triviality and Qualitative Additivity. The Sure-Thing-Principle is a strong axiom which rules out the Allais-preferences discussed earlier and plays a crucial role in establishing the existence of an expected utility representation. The key observation is that we can replace the Sure-Thing-Principle by the much weaker rationality axioms **Certain Prize** and **Alternative Prize** and still show that the comparative probability ordering is a qualitative probability.[26]

Krantz et al. (1971, pp. 208-11) prove a similar result.[27] But instead of

[26]Machina and Schmeidler (1992) also weaken Savage's axiom to give a 'more robust definition of subjective probability'. However, their axioms are stronger than the ones given here, as they entail that preferences always respect stochastic dominance—a property they refer to as 'probabilistic sophistication'—and that preferences are complete. My representation theorem shows how to define subjective probability *without* probabilistic sophistication (and without completeness). Elliott (2017) provides a representation theorem for 'frequently irrational' agents and uses a restricted class of two-outcome acts to construct a unique credence and utility function. However, this credence function is not necessarily a probability function, so this approach does not provide foundations for ascribing subjective *probability* to partly irrational agents.

[27]**Outcome Independence** is equivalent to the first axiom by Krantz et al. (1971), **Certain Prize** is equivalent to their second axiom and **Alternative Prize** is equivalent to their third axiom. They also mention **Non-Degeneracy**.

Restricted Ordering, they assume that preferences are complete and transitive, which rules out incomparable options. Furthermore, I have shown that my axioms are not only sufficient but necessary for the comparative probability ordering to be a qualitative probability. So my result is a strengthening of Krantz et al. (1971), maximally paring down the axioms on the preference relation required to show that the comparative probability ordering is a qualitative probability.

One could also axiomatize comparative probability directly and argue that the qualitative probability axioms are reasonable constraints on belief without trying to justify them by more fundamental axioms about preferences (Joyce 1999, p. 91). However, my project is to show how we can infer credences from preferences without already assuming that we have access to comparative probability judgments. Therefore, I start with axioms on the preference relation.

So far, we have seen how preference reveals qualitative probability. How do we get from qualitative probability to quantitative probability? Probability function p *represents* the qualitative probability $\succcurlyeq$ if for all $X, Y \in \mathscr{F}$,

$$p(X) \geq p(Y) \iff X \succcurlyeq Y.$$

The axioms introduced so far are necessary but not sufficient for the existence of a probability function representing our qualitative probability (Kraft, Pratt, and Seidenberg 1959). To get around this problem, I add an axiom which ensures that the space of events is sufficiently rich to pin down a (unique) probability function. Here is Savage's proposal:

> **Event Richness**. For all $X, Z \in \mathscr{F}$ and outcomes $b, w \in \mathscr{O}$ with $b \succ w$, if $\{b, X; w, X^C\} \succ \{b, Z; w, Z^C\}$, there is a finite partition $\mathcal{Y} = \{Y_1, ..., Y_n\}$ of Ω such that for all $Y_i \in \mathcal{Y}$, $\{b, X; w, X^C\} \succ \{b, (Z \cup Y_i); w, (Z \cup Y_i)^C\}$.

This axiom says that we can cut up events very finely. If you strictly prefer the good prize on X rather than Z, there is a finite partition of our state space such that you still prefer the good prize on X rather than Z or one of the elements of our partition. It is instructive to state **Event Richness** in terms of comparative probability. In these terms, it says that if $X \succ Z$, then there exists a finite partition $\mathcal{Y} = \{Y_1, ..., Y_n\}$ of Ω such that for all $Y_i \in \mathcal{Y}$, $X \succ Z \cup Y_i$.

Why accept **Event Richness**? Savage (1972, p. 38) gives the following

argument. Suppose you judge X to be more probable than Z. Savage points out that we could plausibly choose a coin and throw it sufficiently often such that you would still judge X to be more probable than Z or any particular sequence of heads and tails. As Savage notes, this doesn't require that you consider the coin to be fair. The possible outcomes of the coin flip form the required partition.

Let us end by briefly reflecting on the plausibility of **Event Richness**. Does rationality require that you cut up events very finely? Despite Savage's argument, this does not seem very plausible. Like **Non-Degeneracy**, we should think about **Event Richness** not as a rationality axiom but rather as a structure axiom which ensures that preferences are rich enough to fix subjective probability. This means that realism can tell a more plausible story about this axiom than constructivism. According to the realist story, it is not the existence of subjective probability which requires such a rich event space. Rather, the rich event space is necessary to infer (precise) probability from preference.

Event Richness implies that Ω is infinite.[28] This might strike you as problematic because it seems possible to have subjective probabilities with a finite state space. One option is to look for another structure axiom which is compatible with finite state spaces but still allows us to derive a unique probability function. As Luce (1967) and Fishburn (1986) point out, there are such axioms, but they are rather complicated and do not have intuitive plausibility of **Event Richness**. Since we need some structure axiom anyways, it seems best to stick with **Event Richness** because of its intuitive plausibility. However, finding a good replacement for **Event Richness** which is compatible with finite state spaces is a way in which the representation theorem could be improved.[29]

[28]Proof sketch: Assume Ω is finite. Then consider the least probable event X such that $X \succ \varnothing$. **Event Richness** demands that there exists a finite partition $\mathcal{Y}$ of Ω such that for all $Y_i \in \mathcal{Y}$, $\{b, X; w, X^C\} \succ \{b, Y_i; w, Y_i^C\}$, so $X \succ Y_i \succ \varnothing$, which contradicts our assumption.

[29]Another option would be axioms ensuring that the comparative ordering can be represented by some probability function which needn't be unique (Scott 1964). However, this is not compatible with providing decision-theoretic foundations for *precise* subjective probability and so I will set it aside. One could also argue that comparative probability orderings on finite spaces should be *extendable* to orderings on infinite state spaces which satisfy **Event Richness**.

2.4.3 Representation Theorem

The axioms allow us to prove:

Theorem 2. *If the preference relation $\succsim$ satisfies **Outcome Independence**, **Non-Degeneracy**, **Restricted Ordering**, **Certain Prize**, **Alternative Prize** and **Event Richness**, there is a unique finitely additive probability function $p : \mathscr{F} \to [0, 1]$ representing the comparative probability ordering $\succcurlyeq$, so for all $X, Y \in \mathscr{F}$,*

$$p(X) \geq p(Y) \iff X \succcurlyeq Y.$$

Once we have shown that the comparative probability ordering is a qualitative probability, the rest of the proof is due to Savage. Here is a quick proof sketch inspired by Kreps (1988, pp. 120-125):

Proof. The axioms entail that for any $n \in \mathbb{N}$, there is a partition $\mathcal{Y}$ of Ω into n equiprobable events: events such that $Y_i \approx Y_j$ for each $Y_i, Y_j \in \mathcal{Y}$.[30] We write $C(k, n)$ for a union of k cells of this partition. We define, for any $X \in \mathscr{F}$:

$$k(X, n) = \max_k\big(X \succcurlyeq C(k, n)\big).$$

So given a n-fold equiprobable partition, $k(X, n)$ is the unique maximal positive integer such that X is at least as probable as the union of k cells of our partition. We define

$$p(X) = \lim_{n\to\infty} \frac{k(X, n)}{n}.$$

One can show that p is a finitely additive probability function which represents $\succcurlyeq$ and that it is unique. □

In this proof, we divide Ω into more and more fine-grained equiprobable partitions. For every such partition, we 'approximate' $p(X)$ by the largest number of cells collectively less probable (according to our comparative probability ordering) than X divided by the number of all cells. Step by step, we get a closer approximation, until we recover the 'true' probability of X in the limit. As a simple analogy, think of approximating the area of a two-dimensional figure by drawing more and more fine-grained grids and

[30] Fishburn (1970, pp. 195-8) provides a detailed reconstruction of this step of the proof. Gaifman and Liu (2018) discuss how it relies on the assumption that the events form a σ-algebra.

counting the number of squares covered by the figure divided by the number of all squares. As the grid gets more and more fine-grained, we approximate the area of our figure more and more closely and we recover the true area in the limit.

We can think of the theorem in two ways. If we are inclined towards constructivism, we can think of it as a definition of subjective probability in terms of preferences. In this case, the fact that Alice's subjective probability of rain tomorrow is .3 is *constituted* by the fact that

$$\lim_{n\to\infty} \frac{k(rain, n)}{n} = .3,$$

and from this perspective, my axioms are conditions under which subjective probability exists. If we are inclined towards realism, we think that there is a probability function encoding Alice's beliefs not defined in terms of her preferences. From a realist point of view, we can interpret the proof as giving an algorithm to *measure* Alice's subjective probability by constructing better and better approximations. From this perspective, my axioms are conditions under which subjective probability can be measured by this algorithm.

2.4.4 Countable Additivity

As it stands, the representation theorem delivers a *finitely additive* probability function which represents our agent's beliefs. This probability function might fail to be *countably additive*.[31] Some decision theorists, for example de Finetti and Savage, have argued that rationality only requires finite additivity and violations of countable additivity are fine. However, there are also reasons to want countable additivity. Most importantly, there are convergence theorems in Bayesian statistics which show that under certain conditions, agents with different priors converge to similar opinions after learning enough shared evidence.[32] Many of these convergence theorems require countable additivity (Elga 2016). So if convergence is a central part of your conception of subjective probability, finitely additive probability is

[31]The probability function $p : \mathscr{F} \to [0, 1]$ is countably additive if for any countable sequence $X_1, X_2, ...$ of pairwise disjoint events in $\mathscr{F}$, $p(\bigcup_{n=1}^{\infty} X_i) = \sum_{n=1}^{\infty} p(X_i)$.

[32]Subjective Bayesians draw on such convergence theorems to argue that, despite different priors, rational agents will agree in the long run (Earman 1992, Ch. 6). Convergence arguments also play an important role in some versions of objective Bayesianism (Neth 2023).

not enough. This is not the place to settle whether arguments for countable additivity are conclusive. The key point is that if countable additivity is desirable, we can add another plausible axiom on preferences to ensure that subjective probabilities are countably additive, building on work by Villegas (1964). Details are in appendix A.

2.5 Problems Solved

I explain how my representation theorem does better than Ramsey's method.

2.5.1 Weak Rationality Assumptions

My axioms do not entail that our agent is an expected utility maximizer. They do not even entail the weaker claim that our agent always respects stochastic dominance. A quick way to see this is that my axioms only constrain preferences over a very restricted set of acts—two-outcome acts where one outcome is strictly preferred—while expected utility maximization and stochastic dominance constrain preferences over all acts. My axioms do not even require preferences over arbitrary acts to be transitive. So the axioms are compatible with Allais-preferences and violations of stochastic dominance.

Further, the axioms allow agents to consider many options incomparable. To be sure, **Non-Degeneracy** requires the existence of at least two comparable outcomes. However, the axioms allow agents to consider *all other* outcomes incomparable. Thus, we can make room for the kind of outcome incomparability discussed above (career as a doctor vs. career as a rock star). Furthermore, we allow agents to consider acts with more than two outcomes incomparable even if all outcomes are comparable, like in the complicated portfolio choice discussed earlier. So, speaking a bit loosely, we allow agents to consider almost all options incomparable.

I still make substantive rationality assumptions. In particular, as discussed above, we can imagine agents which violate **Certain Prize** and **Alternative Prize**. It would be desirable to have even more general foundations for subjective probability. But there is a trade-off between substantive rationality axioms which allow us to measure subjective probability but exclude some agents and weak rationality axioms which include these agents but might make measuring subjective probabilities impossible. In particular,

Theorem 1 shows that my rationality axioms are necessary conditions for the agent's comparative probability judgments to be representable by a probability function. The comparative probability judgments of agents which violate these axioms cannot be represented by any probability function. So if we want to further weaken these axioms, a fundamentally different approach to measuring subjective probability is needed.

2.5.2 Useful Advice

As explained above, my axioms allow agents to have some irrational preferences. Thus, we can give useful advice. We can define or measure the subjective probabilities of partly irrational agents from their preferences over simple acts and use these probabilities to give useful advice for how to choose among more complicated acts.

You might complain that my axioms are too weak because they only constrain preferences over test acts. But this is a feature, not a bug. We can measure or define subjective probability from preferences over test acts and then apply your favorite decision-theoretic norm to give advice for choices among more complicated acts. I remain agnostic on what exactly this advice looks like. Beyond the basic requirement to respect stochastic dominance, different decision theorists will give different advice: some of them will advise you to maximize expected utility, others will advise you to maximize risk-weighted expected utility and so on. Since I allow incomparable options, there is also the question of how to decide when options are incomparable. But any decision theorist needs to know at least your credences to give useful advice.[33] My representation theorem shows how we can measure or define your credences without already presupposing that your preferences are fully rational and so enables the decision theorist to give useful advice.

Here is a simple toy example for how we can give useful advice. There is an urn with some red marbles, some yellow marbles and some black marbles. A marble will be drawn from this urn.[34] We observe that Alice strictly prefers winning one dollar if the marble is red over winning one dollar if the marble is black. So Alice prefers $\{\$1, R; \$0, R^C\}$ over $\{\$1, B; \$0, B^C\}$, where R is the

[33]For decision-theoretic advice to be useful, we also need some way to measure your utility function (Narens and Skyrms 2020) and possibly other functions like your risk function (Neth 2019b).

[34]To satisfy **Event Richness**, let us assume that this is the first draw in an infinite sequence of draws from the urn.

event that the marble is red and B the event that the marble is black. We know that Alice likes more money rather than less, so Alice thinks R is more likely than B. Using **Alternative Prize**, we can infer that Alice must judge $R \cup Y$ to be more likely than $B \cup Y$, where Y is the event that the marble is yellow.

Now suppose Alice faces another choice. The first option pays one dollar if the marble is red, two dollars if the marble is yellow and nothing otherwise: $\{\$1, R; \$2, Y; 0\$, B\}$. The second option pays one dollar if the marble is black, two dollars if the marble is yellow and nothing otherwise: $\{\$1, B; \$2, Y; \$0, R\}$. We can advise Alice that, to avoid (strict) stochastic dominance, she should prefer the first option. This is genuinely useful advice because it is consistent with my axioms that Alice has no preference among these options or even prefers the stochastically dominated option.

2.5.3 No Dependence on Utility

My axioms make minimal demands on the richness of the outcome space. I only require that there are at least two outcomes our agent is not indifferent between. Thus, we can define or measure subjective probabilities of a very 'simple-minded' agent who only distinguishes between the outcome GOOD and the outcome BAD and has an ordinal ranking of these two outcomes. We can disentangle measuring subjective probability from the strong assumptions about value implicit in standard representation theorems.

There is a subtle difference in how my structure axioms compare to Ramsey and Savage. While I do not assume a rich space of outcomes, I do assume a rich space of events, as required by **Event Richness**. Ramsey's original method does not need such a rich space of events.[35] So in terms of structural richness, my method does better than Ramsey's in one way but worse in another way. This means that our axioms are incomparable in terms of their logical strength. However, from a philosophical point of view, I think that Ramsey's and my approach assume a similar amount of structural richness. Ramsey assumes that our agent makes very fine-grained distinctions with respect to the value of outcomes while I assume that our agent makes very fine-grained distinctions with respect to the comparative probability of events. In contrast, my rationality axioms are much weaker than Ramsey's

[35] When reconstructing Ramsey's reasoning, Fishburn (1981, p. 151) writes that Ramsey assumes "a finite state set". As noted above, no finite Ω can satisfy **Event Richness**.

rationality assumptions. Compared with Savage, I make the same structural assumptions about event richness but much weaker rationality assumptions, so we have a strictly more general decision-theoretic foundation for subjective probability than Savage's.

The upshot: I have shown how to define or measure subjective probability with much weaker rationality axioms than standard representation theorems. If we are interested in ascribing precise subjective probability to partly irrational agents, this is definite progress. One might also take this result as illustration of how strong **Event Richness** really is. This axiom is doing the heavy lifting in my construction of subjective probability. On the one hand, this might incline some of us to be skeptical of this structure axiom.[36] On the other hand, nobody has figured out how to derive subjective probability without rich preferences and it is probably impossible to do so. So it is fair to say that we have found *better foundations for subjective probability.*

2.6 Interpretation

I sketch two ways in which my representation theorem sheds light on the interpretation of subjective probability. First, it naturally fits with a view on which comparative probability is more fundamental than numerical probability. Second, it suggests a subjectivist version of the classical interpretation of probability.

2.6.1 Comparativism

Our starting point were comparative probability judgments defined in terms of preferences. I laid down axioms to ensure that this comparative probability ordering is a qualitative probability and an additional structure axiom to ensure that there is a unique probability function which represents this ordering. While we ultimately end up with a unique probability function which represents our agent's beliefs, this approach naturally suggests a picture on which comparative probability is *more fundamental* than numerical

[36] Joyce (1999, p. 98) expresses skepticism about the structure axioms in Savage's representation theorem, although one of Joyce's main targets of completeness which I don't assume.

probability.[37] This is in sharp contrast to Ramsey's approach. For Ramsey, subjective probabilities are ratios of utilities and so they are fundamentally quantitative.

The idea that comparative probability is more fundamental than numerical probability has considerable intuitive appeal. It is more natural to think about which of two events is more likely than to assign numerical probabilities. Furthermore, as I will turn to explain now, taking comparative probability as fundamental allows us to tell a plausible story about the axioms.

My representation theorem naturally fits with a combination of realism and constructivism: realism about comparative probability and constructivism about numerical probability. According to this picture, the comparative probability ordering is psychologically real and not reducible to preferences—rather, preferences serve to measure comparative probability. This is the realist aspect. The advantage of this bit of realism is that we can tell a plausible story about some axioms, in particular **Non-Degeneracy**, which requires our agent not to be indifferent among all outcomes. It is not very plausible to think that the *existence* of comparative probability requires this axiom, but much more plausible to think that *measuring* comparative probability requires this axiom.

However, in contrast to the comparative probability ordering, the probability function constructed in the representation theorem is not psychologically real but only a 'representational device' to talk about the underlying comparative probability ordering. This is the constructivist aspect. The advantage of this bit of constructivism is that we can tell a plausible story about **Event Richness**. It is implausible to think that subjective probability requires the rich event space postulated by this axiom. If we think of comparative probability as fundamental, we can say that agents might have comparative subjective probabilities even if they do not satisfy this axiom.

[37]Comparativism is discussed by Koopman (1940), Fine (1973), Zynda (2000), Hawthorne (2017), Stefánsson (2017), Konek (2019), and Elliott (2022). Of course, the idea of starting with comparative probability is well-known in decision theory (Fishburn 1986). However, it is valuable to make explicit that we can be realists about comparative probability and constructivists about numerical probability, while many decision theorists like Savage are constructivist all the way down. As Holliday and Icard (2013) point out, comparative probability can also shed light on probability operators in natural language and so might help us with puzzles about which inferences with these operators are valid (Yalcin 2010; Neth 2019a).

The axiom describes a condition under which we can *represent* comparative subjective probability by a unique probability function, not a condition under which subjective probability *exists*.

2.6.2 Vindicating the Classical Picture

According to the *classical interpretation of probability* associated with Laplace, we can determine the probability of some event as follows.[38] First, we find a suitable set of 'equally possible' cases. Then, we count the number of cases in which the event occurs and divide this number by the number of all cases. For example, if we want to find out the probability of snake eyes (two 1's) when rolling two fair dice, there are 36 'equally possible' cases and exactly one of these cases is snake eyes, so the probability of snake eyes is $\frac{1}{36}$.

There are many well-known objections to the classical interpretation of probability. First, you might complain that the definition given above is circular. Laplace defines probability in terms of 'equally possible' cases, but it is hard to see what 'equally possible' could mean other than 'equally probable'. Second, the classical interpretation entails that all probabilities are rational, since they are the ratio of two positive integers. But there is nothing incoherent about irrational-valued probabilities.[39] Third, what guarantees that we can always find 'equally possible' cases? They are easy to find in games of chance but much harder to find in real-life situations, where we might try to find the probability that a nuclear power plant will have a catastrophic accident in the next 100 years (Halpern 2003, p. 18).

My representation theorem can be construed as *subjectivist version of the classical interpretation of probability*. Recall how we construct the subjective probability function. To find Alice's subjective probability for rain tomorrow, we find n mutually exclusive and collectively exhaustive events which Alice considers to be equally probable. We write down k, the greatest number of events Alice considers to be collectively less likely than rain tomorrow. This is a bit like counting the number of 'cases' in which it rains tomorrow. We approximate Alice's subjective probability of rain tomorrow by k divided by the total number of cases. We define Alice's subjective probability of rain tomorrow as the limit of this procedure as n goes to infinity.

[38] Gillies (2000, Ch. 2) gives an overview of the classical interpretation and Diaconis and Skyrms (2018, Ch. 1) briefly recount the history of reasoning about probability in terms of 'equally possible' cases.

[39] Hájek (1996) uses this as an argument against (finite) frequentism.

The shift towards the subjective helps with some of the well-known worries for the classical interpretation. First, what does it mean to say that the cases are 'equally possible'? In our picture, it means that they are judged to be equally probable by our subject and we can know that they are so judged by looking at our subject's preferences. Since we have not defined comparative probability in terms of numerical probability but rather directly in terms of preferences, we can sidestep the circularity worry. Second, since we define subjective probability as limit of a sequence of rational numbers, we can have irrational-valued subjective probabilities.

Some worries for our subjectivist Laplacean picture still remain. What guarantees that, for any n, we can find a partition of n mutually exclusive and collectively exhaustive events which our subject considers to be equally likely? In our construction, this follows from **Event Richness**, which ensures that the event space of our subject is sufficiently fine-grained. But is it a requirement for the existence of subjective probability to have such a fine-grained event space? Arguably not. If we accept comparativism, we can reply that the more fundamental comparative subjective probabilities still exist without **Event Richness**, but they may not admit of representation by a unique probability function. From this point of view, Laplace's 'equiprobable cases' highlight a condition under which comparative probability judgments can be represented by a unique probability function.

2.7 Conclusion

Ramsey wants to reduce subjective probability to preference but makes very demanding rationality assumptions—the agent under consideration has perfectly coherent preferences. I have shown how to provide better foundations for subjective probability: axioms which ensure that there is a unique probability function representing our agent's beliefs while leaving room for mistakes.

Let me close by observing that we if we are convinced that my axioms are rationally required, we can also read my representation theorem as an answer to the question: *Why be probabilistically coherent?* Because rationality requires you to obey the axioms and if you obey the axioms, there is a unique probability function representing your comparative probability judgments. In contrast to other decision-theoretic arguments for probabilistic coherence, such as dutch book arguments or standard representation theorems, this ar-

gument does not presuppose or entail expected utility maximization. So we have a new argument for probabilistic coherence from weak assumptions about practical rationality in the face of uncertainty.[40]

[40]In contrast to accuracy arguments for probabilistic coherence, we also avoid commitments about epistemic value. Furthermore, accuracy arguments run into difficulties when there are infinitely many possibilities (Kelley and Neth 2023).

Chapter 3

Rational Aversion to Information

We care about learning the truth for its own sake, but we also care about learning because it can lead us to make better decisions. That is, besides the epistemic benefits of finding out the truth, learning often comes with instrumental benefits as well.[1]

Is more information always instrumentally better? Or are there situations in which more information can make us foreseeably worse off? It is clear that information can make us worse off if we consider the cost of processing and storing the information or the opportunity cost of thinking for too long before acting. Nobody thinks that you have to read all the reviews before buying a new vacuum cleaner or that you should think long before hitting the brakes when a red light comes up. It is also clear that information can make us worse off if it is false, so let me be clear that when I talk about information, I always mean true information.

What if the information is cost-free? For rational agents, is it always instrumentally valuable to accept free information? Good (1967) argues that the answer is 'yes' if we accept the principle of maximizing expected utility. However, Good presupposes that you are certain you will update by conditionalization, which means you are certain your new credences after learning are equal to your old conditional credences given the learned event. There are good reasons to assign positive probability to failures of conditionaliza-

[1] A version of this chapter is forthcoming in *The British Journal for the Philosophy of Science*. See `https://doi.org/10.1086/727772` for the published version.

tion, even for rational agents. I show that if you assign a positive probability to failures of conditionalization, the principle of maximizing expected utility can require you to reject free information. Sometimes, even expected utility maximizers are better off knowing less. Moreover, this offers a vindicating explanation of why people sometimes reject information in real-life examples, such as medical testing.

To be clear, this paper is not about situations in which you actually fail to conditionalize. In all my examples below, we can assume that the agent conditionalizes in the actual world. Rather, this paper is about situations in which you fail to be certain that you will conditionalize. You can fail to be certain that you will conditionalize even if you always conditionalize.

Here is the plan. First, I explain Good's argument. Then, I explain why Good's argument presupposes that you are certain you will update by conditionalization and give reasons to reject this assumption. I show how assigning a positive probability to failures of conditionalization can make it rational to reject free information for expected utility maximizers and sketch how this can explain information aversion in the real world. I finish by explaining how we can generalize the value of information to agents who are uncertain about how they will update.

3.1 Good's Argument

I start by introducing some terminology and explain Good's argument.

3.1.1 Terminology

I use the framework of Savage (1972) to model decision making under uncertainty. We have a set Ω of states, which contains all epistemically possible worlds from the point of view of the agent we are modeling. Events are subsets of Ω and we model the credences of our agent by a probability function.[2] We also have a set $\mathscr{O}$ of outcomes, where outcomes contain everything our agent cares about. We model our agent's preferences over outcomes by a utility function u which maps outcomes to their utilities. *Actions* (or acts) are functions from states to outcomes. I assume actions are causally and

[2]I assume Ω is finite and model credences as finitely additive probability function $p : \mathcal{P}(\Omega) \to [0, 1]$.

probabilistically independent of states.[3]

Given a probability function p and utility function u, the *expected utility* of action f is:

$$\mathbb{E}_p(f) = \sum_{\omega \in \Omega} p(\omega)u(f(\omega)).^4$$

I assume your utility function remains fixed through learning.[5] However, your credences change in response to evidence, so I relativize expected utility to a probability function.

A *choice set* is a set of actions among which our agent makes a decision. I assume that all choice sets are finite. Our agent maximizes expected utility if for every available choice set $\mathcal{S} = \{f_1, ..., f_n\}$, she picks an action $f_i \in \mathcal{S}$ which maximizes expected utility relative to her probability function p and utility function u.

I model learning by an *evidence partition* $\mathcal{E}$ of Ω. This partition contains the events our agent might learn, which are mutually exclusive and collectively exhaustive. We can think of the evidence partition as a question, for example the question whether it is sunny or rainy outside. In this case, the evidence partition contains two cells: the worlds where it is sunny outside and the worlds where it is rainy outside. Since the events in the evidence partition are live possibilities for what our agent might learn, they all have non-zero probability, so $p(E) > 0$ for all $E \in \mathcal{E}$.

When learning event E in the evidence partition, our agent updates her credences to $\mathcal{P}_E$. I allow our agent to be uncertain about how she will update. This means that $\mathcal{P}_E$ is not a particular probability function, but rather a random variable whose values can be different probability functions. (Hence the fancy typeface.) The only constraint I impose is that after learning an event, our agent is certain of that event.[6] I write $p(\cdot \mid E)$ for the credences our agent adopts after learning event $E \in \mathcal{E}$ and updating by conditionalization.[7]

Here is an example of our framework in action. You are at the horse

[3]Adams and Rosenkrantz (1980) and Maher (1990) discuss how Good's argument fails if this assumption is relaxed, in both evidential and causal decision theory.

[4]$p(\omega)$ is shorthand for $p(\{\omega\})$.

[5]I set aside cases in which learning leads you to change your utility function, perhaps in a 'transformative experience' (Paul 2014; Pettigrew 2019).

[6]Let $\Delta(\Omega)$ be the set of all probability functions $p : \mathcal{P}(\Omega) \to [0, 1]$. Formally, $\mathcal{P}_E$ is a function from E to $\Delta(\Omega)$ such that for each $\omega \in E$, $\mathcal{P}_E(\omega)(E) = 1$. For each $\omega \in E$, $\mathcal{P}_E(\omega)$ is a particular probability function.

[7]I use the standard ratio definition: $p(A \mid E) = \frac{P(A \cap E)}{p(E)}$ assuming $p(E) > 0$.

track thinking about which horse to bet on. The states specify which horse will win the race (and other relevant facts) and the actions are different bets you might place. The probability function p encodes your credences about different horses winning and the utility function u models how much you value the outcomes of these bets, for example different amounts of money.

Imagine a charming stranger comes up to you and offers you their opinion on which horse is likely to win. Are you willing to listen? The evidence partition $\mathcal{E}$ contains different opinions the stranger might voice and $\mathcal{P}_E$ models how you expect to update your credences after listening. To be clear, the information you learn is not which horse is likely to win but only what the stranger is saying. The stranger might be lying or clueless.

You need to decide: Do you want to find out what the stranger has to say or would you rather place your bet now? It is not obvious how to answer this question. On the one hand, you might listen to the stranger and ignore what they say if you do not find it helpful, so how could listening harm you? On the other hand, perhaps the stranger is trying to mislead you. In this case, do you trust yourself to listen before placing your bet?

3.1.2 The Argument

Good (1967) thinks you should listen to the stranger before placing your bet. More generally, Good argues that if you are rational, then given any choice set and evidence partition, you are never worse off by first learning the true event in the evidence partition and making your choice afterwards rather than making your choice now.[8] Good does not mean that more information always leads to better decisions. You might get unlucky and learn something misleading. Rather, Good argues that learning cannot make you *foreseeably* worse off.

The idea behind Good's argument can be illustrated by an example. Suppose there is a race between horse A and horse B tomorrow. You have to decide whether (i) to bet on horse A, which means you win \$1 if A wins and lose \$2 otherwise, (ii) to bet on horse B, which means you win \$1 if B

[8]Skyrms (1990) provides a helpful overview and points out that Ramsey (1990) and Savage (1972) give similar arguments. Good (1967) notes that his argument is partly anticipated by Raiffa and Schlaifer (1961, p. 90) and Lindley (1965, p. 66). Hosiasson (1931) discusses similar ideas and cites an unpublished paper by Ramsey as inspiration. Interestingly, the argument does not work when you decide whether *someone else* should learn more information before making a decision (Good 1974).

wins and lose \$2 otherwise, or (iii) to play it safe, which means you won't win or lose anything. You think A and B are equally likely to win, so your best option right now is to play it safe. But you can listen to the (accurate) weather report for tomorrow. You think that A is $\frac{3}{4}$ likely to win if it rains and B is $\frac{3}{4}$ likely to win if the sun shines. We can illustrate your decision problem as shown in figure 3.1, where rectangles stand for decisions you face ('choice nodes') and circles stand for events which might happen ('chance nodes').

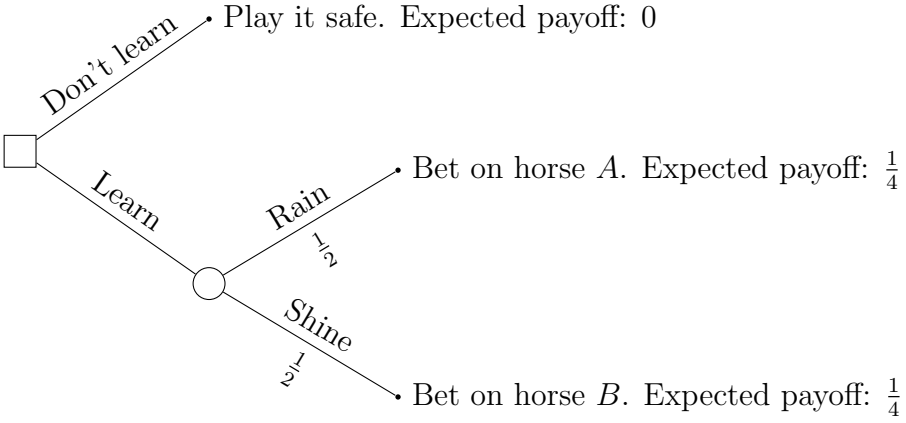

Figure 3.1: If you care about winning, listen to the weather report.

Here is a more general explanation. Good assumes that rational agents maximize expected utility. So, if you are rational, then given some choice set $\mathcal{S}$, you will choose what seems best by your current lights: an action in $\mathcal{S}$ which maximizes expected utility with respect to your current credences. So the expected value of choosing now is the expected utility of one of the best actions in $\mathcal{S}$ relative to your current credences p:

$$\max_{f \in \mathcal{S}} \mathbb{E}_p(f).$$

If, on the other hand, you learn that E is the true event in our evidence partition, you update your credences p to $\mathcal{P}_E$. Good assumes that in each state with non-zero probability, your updated credences are obtained from your current credences by conditionalizing, so $\mathcal{P}_E = p(\cdot \mid E)$. Good also assumes that learning is *cost-free*. This means that before and after learning, you choose among the same actions and outcomes have the same utilities.

The only impact of the information is to change your credences.[9]

After learning, you choose what seems best by your updated lights—an action in $\mathcal{S}$ which maximizes expected utility with respect to your updated credences:

$$\max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f).$$

You don't know yet which element of the evidence partition you will learn. But we can consider the *expected value* of acting after learning:

$$\sum_{E \in \mathcal{E}} p(E) \max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f).$$

Good completes the argument by proving that the expected value of acting after learning is always greater than or equal to the expected value of choosing now:

$$\sum_{E \in \mathcal{E}} p(E) \max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f) \geq \max_{f \in \mathcal{S}} \mathbb{E}_{p}(f).$$

Moreover, this inequality is strict unless there is some action $f \in \mathcal{S}$ which is best regardless of which event in the evidence partition you learn—that is, unless the evidence partition is *irrelevant* for the choice set under consideration. So according to Good, the principle of maximizing expected utility entails:

Value of Learning:

i. Rational agents are always permitted to accept free information before making a decision.

ii. Rational agents are always required to accept free and relevant information before making a decision.

3.1.3 What does the Argument show?

Does Good's argument show that **Value of Learning** is correct? There are ways to push back. One might question the assumption that rational agents

[9]Consider cases in which the information is not free (processing costs, library fees). In such cases, outcomes before and after learning do not have the same utility. Kadane, Schervish, and Seidenfeld (2008, pp. 17-20) discuss this issue in detail. You might also ascribe negative utility to the information itself, for example because it makes you feel bad (Golman, Hagmann, and Loewenstein 2017). I set such cases aside and focus on the instrumental value of information.

always maximize expected utility (Buchak 2010; Campbell-Moore and Salow 2020) or that rationality requires precise credences (Bradley and Steele 2016; Wheeler 2021).[10] Relaxing these assumptions leads to cases where you can be required to reject free information. One might take this to question **Value of Learning**. However, one might also take this as a strike against alternatives to expected utility theory with precise credences.

One might think that the permissibility of accepting free information before making a decision is independently plausible, a piece of common sense: 'look before you leap'. It is a mark in favor of expected utility theory that it entails this piece of common sense and a problem for other decision theories if they conflict with it. From this perspective, Good's argument is not really an argument for **Value of Learning** but rather an argument for expected utility theory. This interpretation is suggested by Kadane, Schervish, and Seidenfeld (2008):

> So the question remains of whether it is reasonable to impose the requirement on a theory of rational decision making that it not require or permit paying not to see cost-free data. If it is, the only such theory known to us is Bayesian decision theory with a single countably-additive proper prior. (Kadane, Schervish, and Seidenfeld 2008, p. 33)

Arguments along these lines are common. The general shape of the argument is that (a) **Value of Learning** is correct and (b) alternatives to expected utility theory are bad because they conflict with this. This assumes that (c) expected utility theory entails **Value of Learning**.[11]

[10] One could also question the assumption that learning can always be modeled as learning an element of a partition (Salow and Ahmed 2019; Dorst 2020; Das 2023), which is foreshadowed by Williamson (2000, pp. 230-7). Merely finitely additive probabilities can also lead to information aversion (Kadane, Schervish, and Seidenfeld 1996; Kadane, Schervish, and Seidenfeld 2008). Since I restrict attention to finite state spaces, the debate over finite versus countable additivity does not concern me here.

[11] More examples: Wakker (1988) shows that violating the independence axiom of expected utility theory leads to situations in which agents reject free information and takes this to show that such violations are irrational. Al-Najjar and Weinstein (2009, p. 249) object to decision theories allowing for ambiguity aversion because they rationalize aversion to information "which most economists would consider absurd or irrational". Briggs (2015) and Ahmed (2016) object to risk-weighted expected utility (REU) theory because it leads to diachronic inconsistency and aversion to information. Buchak (2013, pp. 187–9) also discusses how the value of information can be negative in REU theory and considers this to be a serious cost.

These arguments are misguided. At least, they require serious qualification. This is because expected utility theory, supplemented with plausible assumptions, entails that there are cases in which we are rationally required to reject free information. Good's argument rests on the auxiliary assumption that *you are certain you will update by conditionalization*. This assumption should not be built into expected utility theory and there are good reasons to reject it. If we reject this assumption, expected utility maximizers with precise credences can be required to reject free and relevant information. So (c) is false: expected utility theory does not entail **Value of Learning**. I also argue that (a) is false: rational agents can be required to reject free information. So we should not take **Value of Learning** as axiomatic in our theories of instrumental rationality.

If you are already skeptical of **Value of Learning**, you might argue as follows: expected utility theory entails **Value of Learning** but **Value of Learning** is clearly false. There are many situations in real life where we are better off ignoring free information. Perhaps you think that the stranger at the horse track will try to deceive you. Therefore, we should reject expected utility theory and look for an alternative decision-theoretic framework, perhaps risk-weighted expected utility theory or imprecise credences. I agree that there are many situations in real life where we are better off ignoring free information. However, once we understand that expected utility theory does not entail **Value of Learning**, we can make sense of information aversion within the standard framework of expected utility theory and Bayesian epistemology.

3.2 Against Good's Argument

I explain why Good's argument presupposes that you are certain you will conditionalize (**Immodesty**). I argue that this assumption is implausible. Instead, we should assign some positive probability to not conditionalizing (**Modesty**). I show that expected utility maximization can require modest agents to reject free information.

3.2.1 Good Presupposes Immodesty

As we have seen, Good's argument requires that whichever event in the evidence partition you learn, your future credences are obtained from your

current credences by conditionalization:

The Equation: $\mathcal{P}_E = p(\cdot \mid E)$ for every $E \in \mathcal{E}$.[12]

At first glance, one might think **The Equation** means that you are *actually* a conditionalizer. This is how the assumption is sometimes glossed in presentations of Good's argument.[13] However, **The Equation** actually means that you assign subjective probability one to the event that you will conditionalize.[14] In other words, *you are certain* you will conditionalize. This is because **The Equation** says that for *every* event you might learn, your new credences equal your old credences conditional on the learned event. Taken together, the events in the evidence partition sum to probability one. This means that in every state with positive probability, your new credences equal your old credences conditional on the true event in the evidence partition. Since states represent epistemic possibilities, you are certain you will conditionalize.

So Good's argument presupposes

Immodesty: You are certain you will conditionalize.

Here is another way to bring this out. You might in fact update by conditionalization. Nonetheless, you might assign positive probability to a state in which you fail to conditionalize. In this case, **The Equation** does not hold and Good's argument does not go through. On the other hand, you might be certain you will conditionalize—and so satisfy **The Equation**—but fail to conditionalize in the actual world. This might be because you have assigned probability zero to an unforeseen failure of rationality. In this case, Good's argument still applies. What is at issue is not whether you will actually conditionalize but whether you are certain you will conditionalize. Even if you always conditionalize, you might have good reasons not to be certain of that.

[12]Technically, Good's result requires only that this equality holds with probability one.

[13]For example, Laffont (1989, p. 58) presents a result equivalent to Good's and writes that the agent under consideration "revises his expectations by using Bayes's theorem". This sounds like we're assuming that the agent is actually a conditionalizer.

[14]As Skyrms (1990, p. 247) writes, "the proof implicitly assumes not only that the decision maker is a Bayesian but also that he knows he will act as one. The decision maker believes with probability one that if he performs the experiment he will [...] update by conditionalization [...]". Huttegger (2014, p. 283) also makes this point.

3.2.2 The Case for Modesty

Immodesty is implausible. Instead, we should accept:

> **Modesty:** There is some positive probability that you will not conditionalize.

To be clear, what I have in mind here is subjective probability, not objective chance. So to accept **Modesty** means to assign some positive credence to the possibility that you will not conditionalize.[15]

There are good reasons for **Modesty**. Moreover, these reasons flow from standard principles of Bayesian epistemology. Let me first be clear that it is by no means (physically or metaphysically) necessary that you will conditionalize. Rather, the claim that your new credences after learning are your old credences conditional on the learned event is a substantive claim about how your credences will evolve over time. The following passage by Ramsey makes the point clear:

> [the degree of belief in p given q] is not the same as the degree to which [a subject] would believe p, if he believed q for certain; for knowledge of q might for psychological reasons profoundly alter his whole system of beliefs. (Ramsey 1926, p. 21)[16]

So it is a consistent possibility that you fail to conditionalize. Many Bayesians are attracted to the principle of *regularity*, which says that you should assign positive prior probability to all consistent possibilities. This principle entails **Modesty**.

[15] **Modesty** has been defended before. For example, discussing whether we should defer to our future credences, Briggs (2009, pp. 59–60) writes: "Under all but the most ideal circumstances, agents will have reasons to suspect that future failures of conditionalization are in store". Pettigrew (2020) points out that standard arguments for conditionalization assume 'deterministic updating' and so leave no room for uncertainty about how you will update. Lederman (2015) draws on failures of common knowledge that we will conditionalize to construct counterexamples to Aumann's claim that rational agents cannot 'agree to disagree'. Cohen (2020) discusses uncertainty about updating in the context of epistemic logic. Christensen (2007, p. 3) defends the broader claim that "even an agent who is in fact cognitively perfect might, it would seem, be uncertain of this fact". Similar ideas are defended by many others, including Carr (2019), Bradley (2020), and Dorst (2020).

[16] Diaconis and Zabell (1982) discuss this passage. Of course, the general idea is much more broadly recognized. For example, in *The Portrait of a Lady*, Henry James writes about some piece of news: "But it had been one thing to foresee it mentally, and it was another to behold it actually" (James [1882] 2011, p. 217).

More broadly, that you will conditionalize is an *empirical proposition*. Rationality should not require you to be certain that some empirical proposition is true. For example, if you suffer brain damage as the result of a stroke, you will likely not conditionalize. Plausibly, you should not be certain that you won't suffer brain damage in the future. Therefore, you should not be certain that you will conditionalize.[17]

We can make an even stronger case for **Modesty**. There is a long research tradition in psychology and cognitive science which aims to demonstrate that humans are not ideal Bayesian agents and so do not always conditionalize. There are a number of well-documented fallacies and heuristics which deviate from conditionalization. An example is the *base rate fallacy*, in which people ignore prior probabilities and so overestimate the probability of rare events (Kahneman and Tversky 1973). Another example is the *gambler's fallacy*, which is when people think that a fair coin landing heads provides evidence that the next coin flip will land tails.

Once you learn about these empirical findings, it seems reasonable to believe that they might also apply to *yourself*. Therefore, you should assign some positive credence to not conditionalizing and accept **Modesty**. In addition to such general considerations, you might remember specific cases in which you did not conditionalize, but committed (say) the gambler's fallacy. If you have such evidence, this gives you another strong reason for **Modesty**.

Perhaps you are quite confident of your future rationality. But even if you currently have no evidence that you might fail to conditionalize, it seems reasonable that you might obtain such evidence. For example, you might learn that you just took a drug which increases your susceptibility to the gambler's fallacy or that your brain is wired to misfire in certain situations.[18] Surely, if you learned something like this, you should decrease your credence that your future self will conditionalize. But **Immodesty** rules this out: once you assign probability zero to failures of conditionalization, then no matter what you learn, you will continue to assign probability zero

[17]Note that the possibility of malfunction does not only apply to humans, but also to AI agents and plausibly to any kind of agent which is physically realized.

[18]The reason-impairing drug is inspired by Christensen (2007).

to failures of conditionalization (if you actually conditionalize).[19] This seems unreasonable—surely, there are some things you might learn that would make you doubt your own future rationality. Therefore, you should assign non-zero probability to failures of conditionalization, so **Modesty** follows.

The arguments above appeal to substantive constraints on prior probabilities. Subjective Bayesians reject such constraints beyond adherence to the axioms of probability. So subjective Bayesians will not be moved by my arguments. However, **Immodesty** is also a substantive constraint on prior probabilities and does *not* follow from the axioms of probability. Subjective Bayesians have no reason to accept this constraint.[20]

There are, of course, good reasons to think that rationality requires conditionalization, for example diachronic coherence arguments (Lewis 1999) and various accuracy arguments (Joyce 1998; Greaves and Wallace 2005; Pettigrew 2016). **Modesty** is entirely consistent with this claim. Arguments for conditionalization aim to show:

Conditionalization: You should conditionalize.

Good's argument relies on **Immodesty**, which says that you are certain you will conditionalize. **Conditionalization** does not entail **Immodesty**. We can accept that we should conditionalize but still have good reason to assign positive probability to failures of conditionalization in the future. This is because we might not be certain that our future selves will be rational. Indeed, as good Bayesians, we *should not* be certain that our future selves will be rational if our evidence suggests that we might not be.

Some philosophers have argued that the arguments for conditionalization only support the norm that you should *intend* or *plan* to conditionalize, rather than the norm that you should actually conditionalize.[21] If these philosophers are correct, then it is even harder to see any conflict between

[19]This is the key reason Bayesian epistemologists tend to be skeptical of assigning probability zero to any possible event. For example, Lewis (1980, p. 268) argues that regularity is "required as a condition of reasonableness: one who started out with an irregular credence function (and who then learned from experience by conditionalizing) would stubbornly refuse to believe some propositions no matter what the evidence in their favor."

[20]As Hacking (1967, p. 315) points out, the axioms of probability don't entail that you will actually conditionalize, much less that you are certain of doing so: "The idea of the model of learning is that $Prob(h/e)$ represents one's personal probability after one learns e. But formally, the conditional probability represents no such thing. [...] $Prob(h/e)$ stands merely for the quotient of two probabilities."

[21]This point is discussed, for example, by Pettigrew (2016, pp. 187-88).

the arguments for conditionalization and **Modesty**. We can rationally plan to ϕ while also thinking that there is some positive probability that we will fail to ϕ. For example, I can plan to run a 10K race while also thinking that there is some chance I won't make it to the end.[22]

There are also reasons to doubt whether conditionalization is always rationally required. For example, Douven (2013) argues that an alternative to conditionalization, which he calls 'Inference to the Best Explanation', leads you to converge to the truth faster in some circumstances. If you care about fast convergence, this might be a reason to use Douven's 'Inference to the Best Explanation' instead of conditionalization. While this is no conclusive argument against conditionalization, it might perhaps instill some doubt about whether conditionalization is always rational. Plausibly, the right response to this normative uncertainty is to assign some positive probability to failures of conditionalization even if you are sure you will update rationally.

3.2.3 Modesty entails Information Aversion

Let us assume **Modesty**. I now explain how for modest agents, maximizing expected utility can require you to reject free information. The basic idea is quite simple. If you are modest, you assign some credence to the possibility that learning more information will lead you to make inferences which you do not currently endorse. This might lead you to make choices which, from your current point of view, seem like a bad idea. Therefore, you might be better off ignorant.

Suppose a fair coin will be flipped twice and Ann knows this. She chooses among bets on the second coin flip: a safe bet which always yields zero, a risky bet on heads and a risky bet on tails. Our choice set $\mathcal{S}$ is:

`safe` : {\$0 always},

`risky-heads` : {\$1 if the second coin flip lands heads, −\$2 otherwise},

`risky-tails` : {\$1 if the second coin flip lands tails, −\$2 otherwise}.

Ann values money linearly and is an expected utility maximizer. She is also

[22]Bratman (1992, pp. 11-12) discusses similar examples and argues that one can plan to ϕ without believing that one will ϕ.

certain that her future self will be an expected utility maximizer.[23]

I offer Ann the following choice: She can either make her decision now or learn the outcome of the first coin flip and make her decision afterwards. If Ann makes her decision now, she will pick `safe`. So the expected value of choosing now is:

$$\max_{f \in \mathcal{S}} \mathbb{E}_p(f) = \mathbb{E}_p(\texttt{safe}) = 0.$$

What happens if Ann learns the outcome of the first flip and makes her decision afterwards? If Ann conditionalizes, she will choose the safe bet no matter what she learns since she regards the two coin flips as independent. So there is no reason for her to avoid learning. It can't help her, but it can't hurt her either.

But Ann is modest and assigns some positive probability to failures of conditionalization. In particular, Ann assigns some positive probability to committing the *gambler's fallacy*: after she learns that the first coin flip lands heads, she will become confident that the next coin flip will land tails and vice versa. In particular, Ann assigns some positive probability ϵ to the event that when she learns that the first coin flip lands heads, she will become .9 confident that the second coin flip will land tails and vice versa.

Now suppose Ann learns that the first coin flip lands heads and commits the gambler's fallacy. Relative to her updated credences, the risky bet on tails now looks like the best option. However, given Ann's current credences, the risky bet is the wrong choice. The situation is analogous if Ann learns that the first coin flip lands tails and commits the gambler's fallacy. Figure 3.2 sums up Ann's situation.

The expected value of learning and deciding afterwards is $-\frac{\epsilon}{2}$, *strictly*

[23] I mean that she is certain she will pick one of the actions in $\mathcal{S}$ which maximizes expected utility relative to her updated credences—which might or might not be obtained from her current credences by conditionalization. The value Ann currently assigns to f on the supposition of E is the conditional expected utility $\mathbb{E}_{p(\cdot|E)}(f) = \sum_{\omega \in \Omega} p(\{\omega\} \mid E) f(\omega)$. Gyenis and Rédei (2017) discuss conditional expectations in a much more general setting. The important point is that this conditional expected utility can come apart from the value Ann assigns to f after actually learning E. I also assume that $\mathcal{S}$ does not include actions like 'adopt credence p after learning evidence E'. With such an extended option set, one can argue that certainty that one will maximize expected utility entails certainty that one will conditionalize (Brown 1976), although Pettigrew (2020) points out how uncertainty about updating complicates this argument. Thanks to an anonymous referee for pushing me to clarify how exactly I understand certainty that one maximizes expected utility.

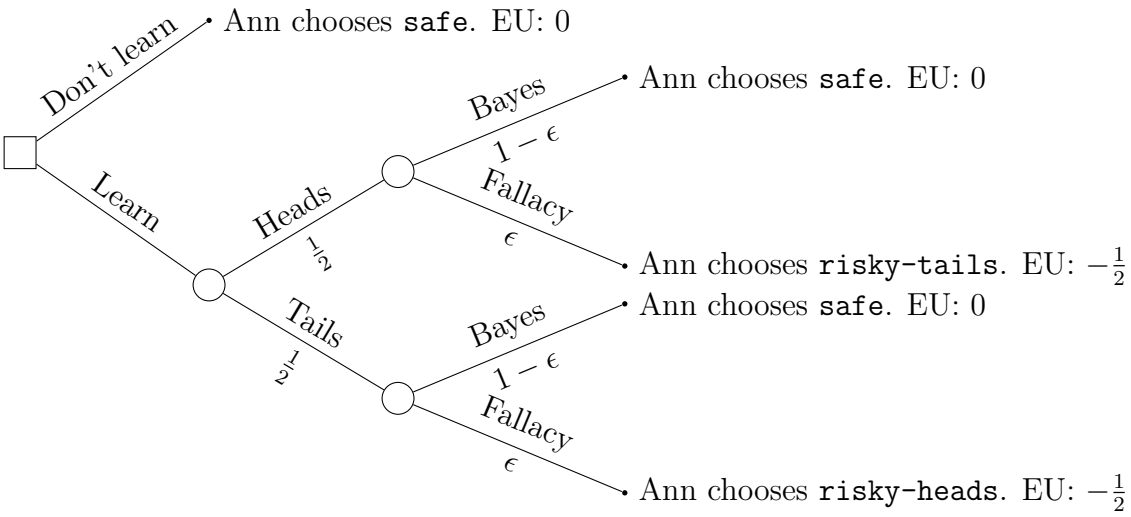

Figure 3.2: Ann's decision problem.

worse than the expected value of choosing now.[24] Learning the outcome of the first coin flip *can* hurt Ann but it can't help her, so she is better off ignorant. Since Ann can foresee all of this, it is rational for her to reject free information. So the principle of maximizing expected utility sometimes recommends rejecting free information.

When I say that the information is 'free', I mean the same as Good: the information does not change the available actions or the utility function. The only impact of the information is to change Ann's credences. And it is not part of the example that Ann ever commits the gambler's fallacy. What makes it rational for Ann to reject the information is not that she actually deviates from conditionalization but that she assigns some positive probability to deviating from conditionalization.

I assume Ann is a 'sophisticated chooser': she predicts her own future choices and takes these predictions into account when making her present

[24]If Ann learns and decides afterwards, she chooses one of the risky options with probability ϵ and the safe option with probability $1-\epsilon$. The risky options have expected utility $-\frac{1}{2}$ and the safe option has expected utility zero. So the expected value of learning and deciding afterwards is $\epsilon \times -\frac{1}{2} + (1-\epsilon) \times 0 = -\frac{\epsilon}{2}$.

decisions (Hammond 1988, pp. 35–6).[25] Since she predicts that her future self might be irrational, she has an incentive to prevent her future self from making bad choices. So Ann faces a predicament similar to Odysseus sailing past the Sirens in Greek mythology. She predicts that learning might compromise her future rationality, so she is better off ignorant.

You might complain that this example is a bit weird. Since Ann regards the two coin flips as independent, there is no way that learning the outcome of the first coin flip could help her make a better choice. At best, the information is neutral. In other words, if Ann is certain she will conditionalize, learning the outcome of the first coin flip is not *relevant* to her choice set. However, we can modify the example so that the information is relevant to her choice set but the principle of maximizing expected utility still recommends rejecting the information.

Again, a coin will be flipped twice and Ann must decide among several bets on the second coin flip. There is a safe bet, a slightly risky bet on heads, a slightly risky bet on tails, a very risky bet on heads and a very risky bet on tails in our choice set $\mathcal{S}$:

`safe` : {\$0 always},

`heads` : {\$1 if the second coin flip lands heads, −\$1 otherwise},

`tails` : {\$1 if the second coin flip lands tails, −\$1 otherwise},

`v-risky-heads` : {\$2 if the second coin flip lands heads, −\$10 otherwise},

`v-risky-tails` : {\$2 if the second coin flip lands tails, −\$10 otherwise}.

This time, Ann does not consider the coin to be fair but thinks that the coin has an unknown bias. The coin might be fair, it might be biased towards heads or it might be biased towards tails—she has no idea. Again, I offer Ann the following choice: She can either make her decision now or learn the outcome of the first coin flip and make her decision afterwards.

Since the coin has an unknown bias, observing the outcome of the first coin flip is informative for Ann. In particular, let us assume that, conditional on the first coin flip landing heads, Ann thinks that the second coin flip lands

[25] Buchak (2013, p. 176) describes the debate around sophisticated choice in decision theory. In moral philosophy, there is a similar debate between actualism and possibilism about what you should do when your future self will act wrongly (Smith 1976; Jackson and Pargetter 1986). Louise (2009) and White (2021) discuss the legitimate role of self-predictions in practical reasoning in more depth.

heads with probability $\frac{2}{3}$. The same goes for tails.[26]

If Ann makes her decision now, she is indifferent between `safe`, `heads` and `tails`. So the expected value of choosing now is:

$$\max_{f \in \mathcal{S}} \mathbb{E}_p(f) = \mathbb{E}_p(\texttt{safe}) = 0.$$

If Ann observes the outcome of the first coin flip, things are more interesting. Suppose Ann will conditionalize. Then if the coin lands heads, Ann will think that the coin is probably biased towards heads, so the slightly risky bet on heads will seem best to her. The very risky bet on heads will still seem too risky. The situation is analogous if the coin lands tails.

But Ann is modest and assigns some positive probability ϵ to the event that she *overweights* the evidence: when she learns that the first coin flip lands heads, she becomes .9 confident that the second coin flip will land heads and vice versa. So Ann takes the evidence into account, but thinks that she might be overconfident in how she does it. There are several reasons for why Ann might do this. She might commit some version of the *base rate fallacy*, ignoring or underweighting prior probabilities. Or she might be susceptible to some form of the *hot hand fallacy*, believing that 'streaks' of successive heads are more likely than warranted by her evidence.

Suppose Ann observes the first coin flip landing heads. If she conditionalizes, she will take the slightly risky bet on heads. But if she is overconfident, she will choose the very risky bet on heads, which looks like a bad choice from her current point of view. A similar story applies if Ann observes the first coin flip landing tails. Figure 3.3 sums up Ann's situation.

The expected value of learning is $\frac{1}{3}(1-\epsilon) - 2\epsilon$.[27] So if Ann thinks the probability of overconfidence is more than $\frac{1}{7}$, the expected value of learning and making her decision afterwards is worse than the expected value of deciding now.[28] So even if learning could be informative, expected utility maximizers can be required to reject free information. Again, it is not part of the example that Ann actually overweights her evidence but only that she

[26]These probabilities can be derived from the 'rule of succession' (Zabell 1989).

[27]If Ann learns and decides afterwards, she chooses one of the very risky options (`v-risky-heads`, `v-risky-tails`) with probability ϵ and one of the less risky options (`heads`, `tails`) with probability $1-\epsilon$. The very risky options have expected utility -2 and the less risky options have expected utility $\frac{1}{3}$. So the expected value of learning and deciding afterwards is $\epsilon \times -2 + (1-\epsilon) \times \frac{1}{3} = \frac{1}{3}(1-\epsilon) - 2\epsilon$.

[28]Choosing now has expected value zero and $0 > \frac{1}{3}(1-\epsilon) - 2\epsilon \iff \epsilon > \frac{1}{7}$.

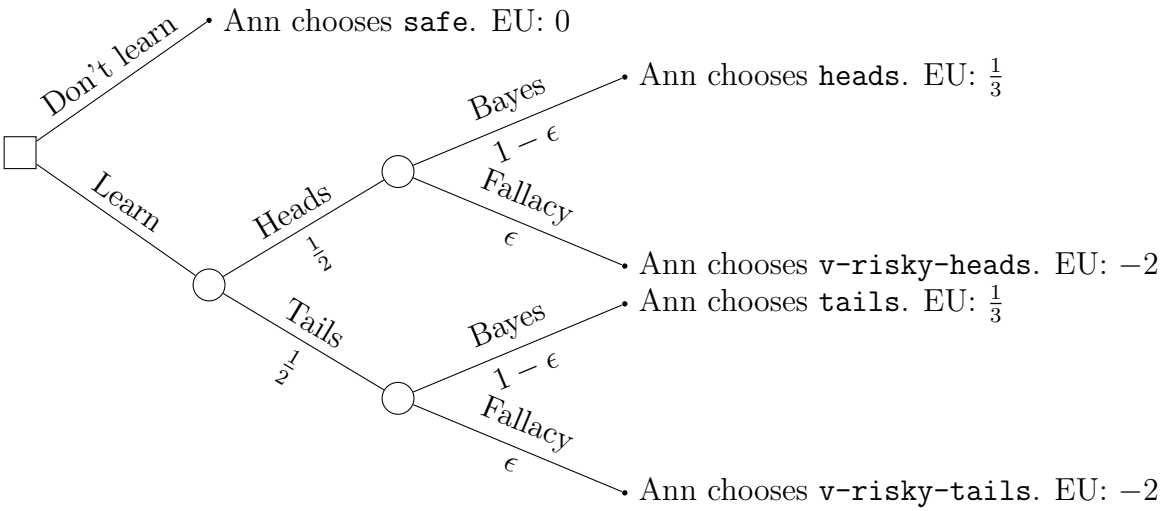

Figure 3.3: Ann's other decision problem.

assigns some probability to doing so.

These example show two things. First, the principle of maximizing expected utility does not imply **Value of Learning**. Expected utility maximizers are not always permitted to accept free and relevant information. Good's argument essentially depends on the assumption of **Immodesty**. If we assume **Modesty**, the principle of maximizing expected utility can require agents to reject free and relevant information.

Second, **Value of Learning** is false: Ann is rational but not permitted to accept free and relevant information. This might seem contentious. Whether or not rationality requires us to always conditionalize, the gambler's fallacy certainly looks irrational. So Ann thinks there is some probability that her future self will be irrational. However, the fact that Ann thinks her future self might be irrational does not entail that Ann is currently irrational. Rationality does not require you to be certain that your future self will be rational.

We can suppose that Ann has good evidence she might commit the gambler's fallacy. All her friends have committed it and she thinks they are relevantly similar to her. In this situation, it is implausible to think that Ann must be certain that her future self will conditionalize. Rather, if she is a good Bayesian, she should take her evidence into account and be modest. We can also suppose that Ann plans to conditionalize. Furthermore, we can

suppose that in the actual word, Ann always manages to follow her plan. She just thinks that there is some chance she might fail. This does not seem like a failure of rationality. Therefore, we should let **Value of Learning** go even if we accept expected utility theory with precise credences and information which partitions logical space.

3.2.4 Information Aversion in the Real World

Moreover, we can use **Modesty** to make sense of real-world cases of information aversion. I will briefly illustrate this with medical testing, blind grading, checking one's stock portfolio and resisting manipulation.

People sometimes reject medical tests. There are several reasons: mistrust of doctors, fear of bad news and so on (Hertwig and Engel 2016, p. 365).[29] **Modesty** suggests another reason. People could fear that the test results might lead them (or their doctors) to draw inferences which they do not currently endorse, resulting in unnecessary treatment and further testing. For example, imagine you learn that a certain marker has increased in your blood test since last time but is still in the normal range. Learning this information might lead you to suspect a worrying trend where there are only random fluctuations. As a result, you might want another test soon which is unnecessary.

Blind grading is often considered good practice. Why is it bad to know the student's names? The standard explanation is that blind grading reduces bias. For example, I might give too much weight to the fact that George got an 'A' on the first paper and treat it as better evidence than it is that his current paper deserves a good grade.

It is sometimes suggested that you shouldn't check your stock portfolio daily. One reason is that it might stress you. However, another reason is that you might be tempted to change the allocation of your portfolio in a way that you currently view as a bad idea. This, in turn, can be explained

[29]Information aversion with respect to medical tests is discussed by Osimani (2012), Jouini and Napp (2018) and many others.

by the risk of overweighting the significance of small fluctuations.[30]

It appears rational to refuse information designed to manipulate you. For example, one reason to avoid social media is that the information shown on your feed is designed to influence your behavior: to make you buy the products advertised there, to make you spend even more time on the platform and so on (Véliz 2020, pp. 69–76). Even if we sidestep issues of misinformation and assume the information you see on your feed is accurate, the fact that this information is designed to influence your behavior by companies which do not have your best interest at heart is a reason to stop looking. A similar case is when you refuse to talk to a manipulative person. Even if everything the manipulative person says is true, you might be better off not listening. This is because you might suspect that you will not update rationally on information designed to manipulate you.

Many other examples of information aversion in real life can be explained along similar lines.[31] There are, of course, competing explanations: perhaps people deviate from expected utility theory, have imprecise credences or assign negative utility to bad news. But in the examples above, it seems independently plausible that we assign some probability to *overweighting evidence*: giving too much weight to the result of medical tests, the past performance of our students, small fluctuations in our stock portfolio and the information on our feed. When we reflect on how much weight we should assign to this information, we might conclude: 'a little bit, but not very much'. But once we actually learn the information, we might assign more weight to it than we have previously considered rational. Imagine a positive result on a medical test slightly increases your probability of serious illness. Before doing the test, you might calmly assign conditional probabilities which reflect this slight increase. But when you learn that the test actually turned out positive, you might increase our probability of serious illness more than you previously considered warranted.

If we model overweighting evidence as deviation from conditionalization,

[30] In a best-selling popular science book on computer science and decision theory, Christian and Griffiths (2016, p. 148) write: "If you want to be a good intuitive Bayesian—if you want to naturally make good predictions, without having to think about what kind of prediction rule is appropriate—you need to protect your priors. Counterintuitively, that might mean turning off the news". They do not consider how we can make sense of this idea without contradicting Good's theorem. **Modesty** offers an elegant way of doing so.

[31] For example elite-group ignorance, which Kinney and Bright (2021) explain using risk-weighted expected utility theory. Yong (2023) critically discusses this explanation.

we have seen that it can be a good idea to reject free information. So **Modesty** can explain these examples of information aversion in the real world in a way that seems to get at the heart of the matter. On the other hand, explaining these cases by saying, for example, that people are not expected utility maximizers seems to have less independent motivation. So while I have no conclusive argument that **Modesty** is the correct explanation for these cases, it seems like a particularly plausible candidate.[32]

3.3 Value of Information Generalized

Surely, modest agents are not always required to reject free information. Even if you have some uncertainty about how you will update, this does not mean that you are always better off ignorant. But how should modest agents decide when to learn more information? And how general is the link between information aversion and **Modesty**? I answer these questions by generalizing the value of information to modest agents.

3.3.1 Good's Value of Information

Good's argument gives us a way to measure the value of information. To state this idea, it is useful to introduce an additional bit of notation. I write $\mathcal{P}(\cdot \mid \mathcal{E})$ for your credences updated by conditionalization on the evidence partition $\mathcal{E}$. This is a random variable which takes different probability functions as values in different state.[33] Then, we can define the value of information as follows (Blackwell 1951; Raiffa and Schlaifer 1961; Howard 1966):[34]

Definition 3. *The* value of information *for* $\mathcal{E}$ *is:*

$$Val_{Good}(\mathcal{E}) = \mathbb{E}_p\left(\max_{f \in \mathcal{S}} \mathbb{E}_{\mathcal{P}(\cdot|\mathcal{E})}(f)\right) - \max_{f \in \mathcal{S}} \mathbb{E}_p(f).$$

[32] Thanks to an anonymous referee for pushing me to clarify why **Modesty** is a plausible explanation of these cases of information aversion.

[33] More rigorously, we can define $\mathcal{P}(\cdot \mid \mathcal{E})$ as the random variable which maps each $\omega \in \Omega$ to $p(\cdot \mid E)$ for the unique $E \in \mathcal{E}$ such that $\omega \in E$.

[34] Le Cam (1996) sketches the history of this concept, which apparently goes back to an unpublished RAND memorandum entitled 'Reconnaissance in Game Theory' based on suggestions by von Neumann (Bohnenblust, Shapley, and Sherman 1949).

This is the difference between the expected value of choosing after learning and the expected value of choosing now. It measures how much you expect the information to improve your decision. In this context, **Value of Learning** is captured by the fact that for any evidence partition $\mathcal{E}$, $Val_{Good}(\mathcal{E}) \geq 0$. In slogan form: 'the value of information is always non-negative'.

This concept is useful because it allows us to say *how much* you should value learning the answer to a question. It also allows us to compare the value of learning the answers to different questions. This means we can formalize tradeoffs between acting now versus learning more and acting later even if learning is costly. Such tradeoffs are ubiquitous. In many real-life contexts, such as drug trials, we have to decide how much to sacrifice for learning more information.[35] So it is not surprising that the value of information is widely used in economics and artificial intelligence.[36] However, the standard way of defining the value of information presupposes **Immodesty**.

3.3.2 General Value of Information

Here is a proposal for how we can define the value of information in a more general way. I write $\mathcal{P}_{\mathcal{E}}$ for your credences updated on the evidence partition $\mathcal{E}$ without assuming you are certain you will update by conditionalization. This is a random variable whose values are different probability functions in different states.[37]

Suppose you will learn the true element of some evidence partition $\mathcal{E}$. Then you update your credences in some way—perhaps you conditionalize, perhaps you do something different—and choose the action which maximizes

[35]Such decision problems can be formalized as 'multi-armed bandits' in which one must balance exploiting, acting according to one's current best estimate, and exploring new and potentially better options (Lattimore and Szepesvári 2020). The value of information can be used to define optimal solutions to such problems.

[36]Russell and Norvig (2018, pp. 628-33) discuss the value of information in AI research. Hadfield-Menell et al. (2017) discuss a model of how to ensure that AI agents always defer to humans which relies on the value of information being non-negative. In addition, the value of information is relevant to much other work, for example in the philosophy of language (Van Rooy 2003), to discussions about 'longtermism' in ethics (Askell and Neth forthcoming) and the epistemology of disagreement (Dorst forthcoming).

[37]On each $E \in \mathcal{E}$, $\mathcal{P}_{\mathcal{E}}$ agrees with $\mathcal{P}_E$ as defined in section (3.1.1).

expected utility relative to your updated credences:

$$\arg\max_{f \in \mathcal{S}} \mathbb{E}_{\mathcal{P}_\mathcal{E}}(f).^{38}$$

This expression will usually denote different actions in different states, because you might learn different events and update on those events in different ways. I assume that there is a unique best action in every state.[39]

We are interested in evaluating how good this action is from your current perspective, so we consider the expected utility of this action given your current credences:

$$\mathbb{E}_p\left(\arg\max_{f \in \mathcal{S}} \mathbb{E}_{\mathcal{P}_\mathcal{E}}(f)\right).$$

This is the expected utility of the action you think you will actually do after learning. We model a 'sophisticated chooser': our agent predicts her future choices and takes this information into account when making present decisions. I propose the following definition:

Definition 4. *The* general value of information *for* $\mathcal{E}$ *is:*

$$Val_{General}(\mathcal{E}) = \mathbb{E}_p\left(\arg\max_{f \in \mathcal{S}} \mathbb{E}_{\mathcal{P}_\mathcal{E}}(f)\right) - \max_{f \in \mathcal{S}} \mathbb{E}_p(f).$$

This measures the difference between the expected utility of your current best action and the expected utility of the action that you think you will choose after learning. In contrast to Good, I do not assume that you are certain you will conditionalize. I still assume you are certain you will maximize expected utility.

If we assume **Immodesty**, my proposal is identical to Good's:

Theorem 4. *If* $\mathcal{P}_E = p(\cdot \mid E)$ *for all* $E \in \mathcal{E}$*, then* $Val_{General}(\mathcal{E}) = Val_{Good}(\mathcal{E})$.

[38] The term $\arg\max_{x \in X} g(x)$ denotes the argument of the maximum: the $x \in X$ such that $g(x)$ is maximal.

[39] So for every $\omega \in \Omega$, there is a unique $f^* \in \mathcal{S}$ maximizing $\mathbb{E}_{\mathcal{P}_\mathcal{E}(\omega)}(\cdot)$. Recall that $\mathcal{P}_\mathcal{E}$ is a function from states to probability functions, so $\mathcal{P}_\mathcal{E}(\omega)$ is a particular probability function—the credence you adopt after learning the true element of $\mathcal{E}$ in state ω. I do not consider cases where two actions are tied for the best action because in such cases, we would need to consider how to break the tie (introduce a selection function), which leads to additional complications. Buchak (2013, p. 190) provides relevant discussion and references on how indifference complicates sophisticated choice.

The proof is in appendix B. In contrast to Good's value of information, the general value of information can be negative. We have seen this in the examples above. But it is not always negative. This is shown by the second example above. When Ann considers the possibility of overconfidence sufficiently unlikely, she is better off observing the first coin flip. In (slightly clunky) slogan form: 'the value of information is sometimes negative, but not always. It depends'.

We can also say something about how general the link between **Modesty** and information aversion is. For this purpose, I make two additional assumptions. *Utility Richness* says that for every $x \in [0,1]$, there is some outcome $o \in \mathscr{O}$ such that $u(o) = x$.[40] *Evidential Independence* says that conditional on the learned event, your updated credences are independent of what action is best.[41] Evidential Independence rules out cases where you deviate from conditionalizing because you become more certain of the truth. For example, you might observe that a fair coin lands heads and be able to foresee that it lands tails next. If we agree that your evidence is that the coin lands heads on the first flip, you do not update by conditionalizing on your evidence. However, clairvoyance can lead you to make better decisions than conditionalization. In contrast, I consider deviations from conditionalization which are not systematically correlated with which action is actually best. I have implicitly made this assumption earlier: Ann is equally likely to commit the gambler's fallacy whether the second coin flip lands heads or tails.

Assuming Evidential Independence, we can write $Val_{General}(\mathcal{E})$ as:

$$Val_{General}(\mathcal{E}) = \sum_{E \in \mathcal{E}} p(E) \sum_{i=1}^{n} p(\text{choose } f_i \mid E) \mathbb{E}_{p(\cdot|E)}(f_i) - \max_{f \in \mathcal{S}} \mathbb{E}_p(f),$$

where 'choose f_i' denotes the event that you choose action f_i after learning

[40] Our outcome space could contain lotteries which yield outcome b with probability p and outcome w with probability $(1-p)$. In this case, we only need two 'primitive' outcomes b and w with $u(b) > u(w)$ to obtain rich utilities.

[41] More precisely, for every $E \in \mathcal{E}$, your updated credences $\mathcal{P}_E$, which determine which action you will choose after learning, are independent of all $f \in \mathcal{S}$ conditional on E. This means, in particular, that for all $f, g \in S$, $\mathbb{E}_{p(\cdot|E \cap \text{choose } f)}(g) = \mathbb{E}_{p(\cdot|E)}(g)$, where 'choose f' is the event that you choose action f after learning. The intuition is that learning that you choose a particular action after learning E does not affect the expected utility of actions beyond learning E. It would be interesting to investigate cases where deviations from conditionalization are systematically correlated with which actions are best. Here, I focus on the simple case where Evidential Independence holds.

E, which means that f_i maximizes expected utility relative to your updated credences after learning E. Note that $p(\text{choose } f_i \mid E)$ is your current conditional probability that you will choose action f_i after learning E. You evaluate how good this action is by its conditional expected utility $\mathbb{E}_{p(\cdot|E)}(f_i)$ given your current credences. If you do not conditionalize, this conditional expected utility can come apart from the unconditional expected utility of the action according to your updated credences.[42]

We can show the following:

Theorem 5. *Assuming Utility Richness and Evidential Independence, for every modest agent, there is some choice set where* $Val_{General}(\mathcal{E}) < 0$.

The proof is in appendix B. Given our assumptions, any positive probability of not conditionalizing leads to information aversion. It does not matter why we are modest, as long as Evidential Independence holds. The examples above demonstrated how *psychological uncertainty* about whether you will update rationally leads to information aversion. The theorem shows that even if you are certain that your future self will be rational, *normative uncertainty* about whether conditionalization is rational leads to information aversion. So we have a very general argument against **Value of Learning**. This also means that we cannot rescue Good's argument by saying that while we might not be certain we will conditionalize, we are very confident we will conditionalize. (Perhaps we have a 'default entitlement' to believe in our future rationality.) Any non-zero probability of failing to conditionalize means trouble for Good.

We can also show:

Theorem 6. *Assuming Evidential Independence,* $\mathcal{E}$, $Val_{General}(\mathcal{E}) \leq Val_{Good}(\mathcal{E})$ *for every evidence partition* $\mathcal{E}$.

The proof is in appendix B. Doubts about how you will update cannot increase the value of information. Note that one might take this theorem as a reason to think that you should conditionalize, at least relative to the assumption of Evidential Independence. But we are often not sure whether we will be rational in the future and cannot change anything about that. If we are in such a situation, the theorem tells us that we should value learning less than if we were certain that we would conditionalize.

[42] Thanks to an anonymous referee for suggesting to make the formula for $Val_{General}(\mathcal{E})$ more explicit and see Lemma (1) in Appendix B.

3.4 Conclusion

Good argues that the principle of maximizing expected utility entails **Value of Learning**: rational agents are always permitted to accept free information and required to accept information which is free and relevant. I have argued that the principle of maximizing expected utility does not entail **Value of Learning** and that **Value of Learning** is false. The key observation is that Good's argument only works if we are certain that we will update by conditionalization but we have good reason not to be.

What follows? First, we can give better advice to modest agents: sometimes, they are better off ignorant. Since we arguably are—and should be—modest, this advice applies to us. Sometimes, we are better off ignorant. Sometimes, we should avert our eyes and stuff our ears with wax to avoid learning the song of the Sirens. Second, proponents of expected utility theory should be careful when objecting to alternative decision-theoretic frameworks on the grounds that these frameworks sometimes permit or require agents to avoid free information. Properly understood, expected utility theory does the same. So this objection loses much of its dialectical force. Third, information aversion is a feature and not a bug. Plausible arguments from Bayesian epistemology push us towards **Modesty**. And once we accept **Modesty**, we can explain many instances of information aversion in the real world which would otherwise be puzzling. By going beyond Good, we end up with a better decision theory.

Chapter 4

Against Coherence

I'll say that an agent is *(diachronically) incoherent* if they are disposed to make a sequence of choices over time which leads to sure loss. Then, we have the following principle:

Coherence. It is not rationally permissible to be incoherent.

Many philosophers accept **Coherence** and draw on it to argue for other rationality constraints. For example, Lewis (1999) uses **Coherence** to argue that rationality requires us to update by conditionalization. Other philosophers use **Coherence** to argue against non-standard decision theories such as risk-weighted expected utility theory (Buchak 2013; Briggs 2015).

I'll say that an agent is *modest* if they are uncertain about how they will update their beliefs after learning some piece of evidence. Then, we have the following principle:

Uncertainty. It is rationally permissible to be modest.

I will show that **Coherence** and **Uncertainty** are in conflict and argue that we should given up **Coherence**. There are two key reasons. First, **Uncertainty** is very plausible and supported by considerations from Bayesian epistemology. Second, the kind of incoherence which arises from **Uncertainty** is analogous to buying insurance, which is rationally permissible. Therefore, **Coherence** must go. I'll close by discussing connections to the reflection principle, which says roughly that you should defer to your future credences. For similar reasons, we should reject the reflection principle.

4.1 Lewis' Diachronic Dutch Book

Lewis (1999) uses **Coherence** to argue that rationality requires us to update by conditionalization. I'll start by explaining Lewis' argument.

Suppose you are about to learn some information. We'll model your situation by an *evidence partition* $\mathcal{E}$ on a finite set of states Ω. We'll assume that your prior credences can be modeled by a (finitely additive) probability function $p : \mathcal{P}(\Omega) \to [0, 1]$. Since $\mathcal{E}$ represents live possibilities for what you might learn, we'll assume $p(E) > 0$ for all $E \in \mathcal{E}$.[1]

An *update policy* π specifies how you plan to respond to each piece of evidence you might learn. For each $E \in \mathcal{E}$, $\pi(E)$ is the credence you plan adopt after learning evidence E. More precisely, an update policy is a function $\pi : \mathcal{E} \to \Delta(\Omega)$, where $\Delta(\Omega)$ is the set of all probability functions on $\mathcal{P}(\Omega)$. *Conditionalization* is the update policy which says that for each $E \in \mathcal{E}$, you plan to update by conditionalizing your prior p on E, so $\pi(E) = p(\cdot \mid E)$.[2]

For our purposes, a *bet* is a function $b : \Omega \to \mathbb{R}$ from states to payoffs in your favorite currency. For each state $\omega \in \Omega$, $b(\omega)$ is the payoff you get when holding bet b in state ω. The *fair price* of bet b is its expected value relative to your prior, which is $\sum_{\omega \in \Omega} b(\omega)p(\omega)$.[3] We'll assume that agents are always willing to buy and sell bets for their fair price. This is equivalent to assuming that our agent is an expected utility maximizer with linear utility function.[4] A bet is *fair* if its expected value is non-negative. A bit more precisely: a bet offered before updating is fair if its expected value relative to p is non-negative and a bet offered after updating is fair if its expected value relative to $\pi(E)$ is non-negative, where E is the learned element of evidence partition $\mathcal{E}$.

A *diachronic dutch book* is a sequence of bets offered before and after learning such that you (your prior and update policy) consider each individ-

[1] I set aside learning events with prior probability zero (Rescorla 2018; Meehan and Zhang forthcoming).

[2] I'll use the standard ratio definition: $p(A \mid E) = \frac{p(A \cap E)}{p(E)}$ whenever $p(E) > 0$.

[3] $p(\omega)$ is shorthand for $p(\{\omega\})$.

[4] The assumption of linear utility can be relaxed by replacing monetary payoffs with payoffs in utils, but the assumption of expected utility maximization cannot be relaxed. Buchak (2013, pp. 201–12) discusses dutch book arguments from the perspective of an alternative decision theory.

ual bet to be fair, but taken together, the bets lead to sure loss.[5] So if there is a diachronic dutch book against you, you are incoherent. Lewis (1999) shows that if π is any update policy other that conditionalization, there is a diachronic dutch book against π.[6] Therefore, violations of conditionalization lead to incoherence. Moreover, there is no diachronic dutch book against conditionalization (Skyrms 1987). So assuming **Coherence**, we have a powerful argument for why rationality requires conditionalization.

Here is an example. Aggu is about to observe two flips of a coin he believes to be fair. Aggu does not conditionalize. Instead, after observing that the first coin flip lands heads, Aggu becomes .9 confident that the second coin flip will land tails. So Aggu commits the gambler's fallacy when observing that the first coin flip lands heads. He thinks that tails is 'due' now.

Here is a diachronic dutch book against Aggu. Consider the following bets with fair prices relative to Aggu's prior credences:

- A pays 1 if heads twice. Fair price: $\frac{1}{4}$.
- B pays $\frac{1}{2}$ if tails first. Fair price: $\frac{1}{4}$.
- C pays $\frac{2}{5}$ if heads first. Fair price: $\frac{1}{5}$.
- D pays 1 if heads second.

All bets pay zero otherwise. I omit the fair price for D since it is offered after learning.

As mentioned above, I assume that Aggu is always willing to buy and sell bets from us for their fair prices. (We have the honorable position of the bookie.) If we *sell* a bet to Aggu, he pays us the fair price of the bet and we have to pay out whatever the bet turns out to be worth. If we *buy* a bet from Aggu, we pay Aggu the fair price of the bet and he has to pay us whatever the bet turns out to be worth.

[5] In contrast, a *synchronic dutch book* is a sequence of bets offered at a single time such that you consider each individual bet to be fair, but taken together, the bets lead to sure loss. If your credences violate the axioms of probability, you are subject to a synchronic dutch book (Ramsey 1926; de Finetti 1937). If your credences are probabilistic, you are immune to synchronic dutch books. Since I'll assume that credences are probabilistic, all agents discussed below will be immune to synchronic dutch books.

[6] Lewis' argument was first reported by Teller (1973). Freedman and Purves (1969) prove a similar result.

Now our diachronic dutch book against Aggu works as shown in figure 4.1. We start by selling bets A, B and C to Aggu for their fair prices. If the first coin flip lands tails, we are done. This is because we have to pay Aggu for bet B but since we also sold A and C, Aggu suffers sure loss. If the first coin flip lands heads and Aggu commits the gambler's fallacy, we buy bet D from Aggu for $\frac{1}{10}$. As a result, Aggu makes a net loss in every possible state. Moreover, even though Aggu can see ahead of time that he will suffer sure loss, each individual bet seems fair to him. This seems to indicate some kind of inconsistency or irrationality on Aggu's part.

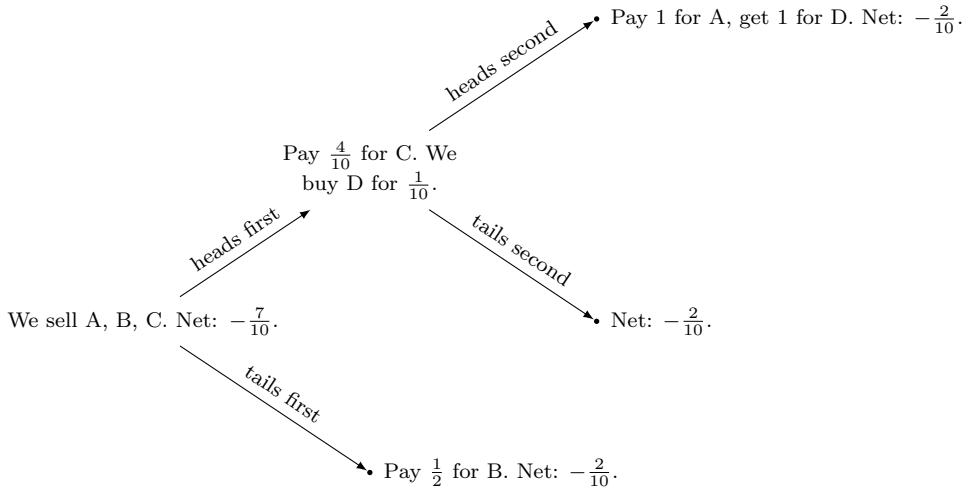

Figure 4.1: Diachronic dutch book against Aggu.

Since the bets are fair, it is permissible for Aggu to accept them but also to reject them. As Skyrms (1993, p. 320) observes, we can strengthen our diachronic dutch book by making the bets better than fair. We take a fraction of our sure winnings and use that to 'sweeten' each of the bets. Then, each individual transaction will seem better than fair to Aggu, a deal so good he cannot refuse. But taken together, Aggu will still suffer sure loss.

By tweaking the payoffs of the bets, we can make sure that Aggu suffers sure loss for *any* deviation from conditionalization.[7] Furthermore, by in-

[7]In particular, we can tweak the payoff of bet C. In general, C pays δ if heads-first, where E is heads-first and $\delta = p(\text{heads-second} \mid E) - \pi(E)(\text{heads-second})$. This assumes that Aggu assigns less credence to heads second than required by conditionalization (δ is positive). If Aggu assigns more credence, we first buy A, B and C, then sell D.

creasing the payoffs of the bets, we can make Aggu's losses arbitrarily high. Note how Aggu would avoid the diachronic dutch book by conditionalizing. If Aggu would conditionalize on his evidence, he wouldn't be willing to buy D for $\frac{1}{10}$ after observing that the first coin flip lands heads. But since he comes to be very confident that the next coin flip will land tails after observing heads, he considers $\frac{1}{10}$ to be a fair price for bet D. Sure loss results.

There are diachronic dutch book arguments for other norms besides conditionalization. For example, there are diachronic dutch book arguments (or 'money pumps') against agents with cyclic preferences (Davidson, McKinsey, and Suppes 1955). More generally, philosophers have objected to alternatives to expected utility theory because they can lead agents into incoherence (Machina 1989; Steele 2010; Buchak 2013; Briggs 2015; Gustafsson 2022). There are diachronic dutch book arguments against various solutions to the sleeping beauty problem (Hitchcock 2004), imprecise credences (Elga 2010), causal decision theory (Oesterheld and Conitzer 2021) and violations of the reflection principle, which says roughly that you should defer to your future credences (van Fraassen 1984). (I return to reflection below.)

What do diachronic dutch book arguments show? Some object that dutch book arguments are too pragmatic and so cannot ground epistemic norms like conditionalization.[8] I set these concerns aside. Others object to assumptions built into the set-up, for example the assumption that the pieces of evidence you might learn form a partition (Gallow 2019; Das 2020). I will accept the assumptions about the structure of evidence, but point out that Lewis' diachronic dutch book argument makes another assumption: you are never uncertain about how you will update your credences after learning a given piece of evidence. I will show that we can generalize diachronic dutch books to agents who are uncertain about updating, but argue that this is bad news for diachronic dutch book arguments.

4.2 A New Diachronic Dutch Book

In Lewis' set-up, updating is deterministic. For every piece of evidence you might learn, your updating policy specifies a unique credence you'll adopt after learning. But what if you are uncertain about how you will react to some piece of evidence? There are often good reasons for such uncertainty.

[8]This is a key motivation for alternative accuracy-based justifications for epistemic norms like conditionalization (Joyce 1998; Easwaran 2013; Pettigrew 2016).

For example, you might suspect that there is some chance you'll succumb to the gambler's fallacy even though you plan to conditionalize. Perhaps you have seen your friends in the grip of the gambler's fallacy. What happens if we admit such uncertainty?

It turns out that if you are uncertain about how you will update after learning, you are subject to a diachronic dutch book. Here is an example. Like Aggu, Beatrice is about to observe two flips of a coin she believes to be fair. Unlike Aggu, Beatrice just thinks that there is a 10% chance she'll commit the gambler's fallacy. Otherwise she'll conditionalize. (Or she does something different—it won't matter for the argument.)[9]

Before getting into the details, here is a heuristic argument for why we can build a diachronic dutch book against Beatrice. Either Beatrice will violate conditionalization or she won't. If she violates conditionalization, we can use Lewis' diachronic dutch book to inflict sure loss upon her. So the only question is: how can we make sure Beatrice suffers sure loss even if she conditionalizes? We can do so by selling Beatrice *insurance against not conditionalizing*: bets which pay off if she does not conditionalize. If she conditionalizes, she paid for the insurance but got nothing back, so she suffers sure loss. Now we just have to make sure that if Beatrice violates conditionalization, she still suffers sure loss despite the insurance payout. The result is that Beatrice suffers sure loss by her own lights, even if she actually conditionalizes.

Here are the details. Consider the following bets with fair prices relative to Beatrice's prior credences:

- A pays 1 if heads twice and she doesn't conditionalize. Fair price: $\frac{1}{40}$.
- B pays $\frac{1}{2}$ if tails first or (heads first and she conditionalizes). Fair price: $\frac{19}{40}$.
- C pays $\frac{2}{5}$ if heads first and she doesn't conditionalize. Fair price: $\frac{1}{50}$.
- D pays 1 if heads second.

As before, all bets pay zero otherwise and D is offered after learning.

[9] I assume Beatrice is uncertain about her actual updating behavior, which is compatible with her following a deterministic updating policy but not knowing this to be so. A subtly different question, which I set aside, is whether Beatrice can permissibly plan to update in a chancy way (Pettigrew 2020).

Taken together, A and C are insurance against not conditionalizing. We start by selling A, B and C to Beatrice for their fair prices. If the first coin flip lands tails, we are done. If the first coin flip lands heads and Beatrice conditionalizes, we are done as well.[10] She paid for the insurance but got nothing back. If the first coin flip lands heads and Beatrice violates conditionalization, we buy D. As a result, Beatrice suffers sure loss. The details are shown in figure 4.2.

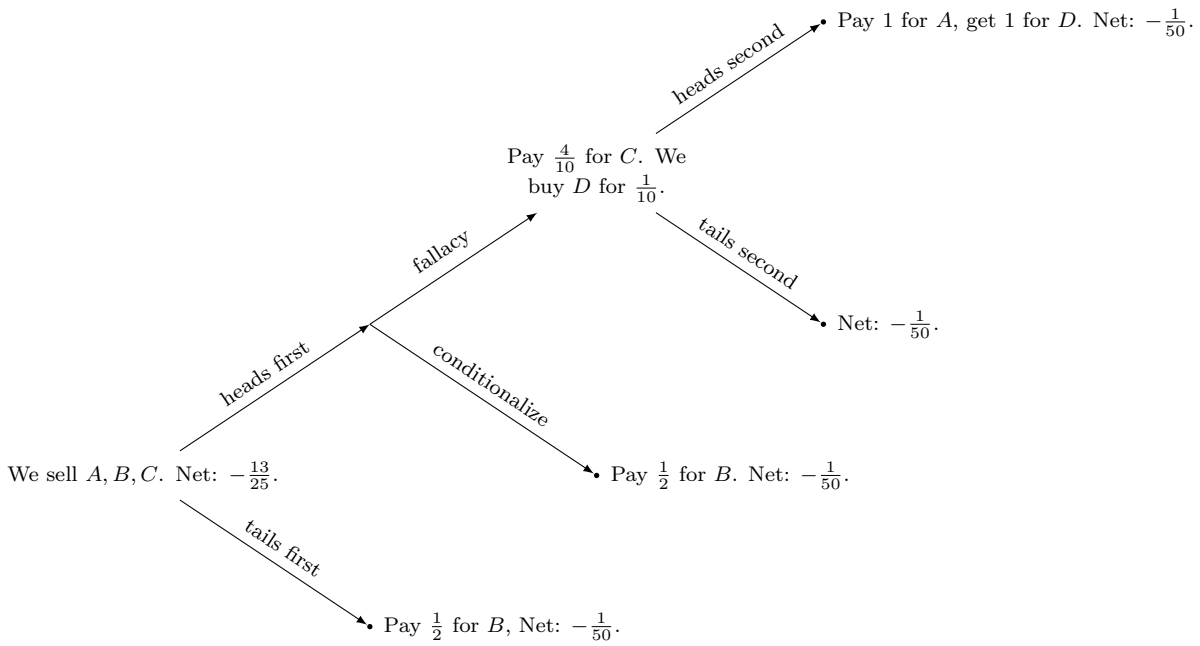

Figure 4.2: Diachronic dutch book against Beatrice.

We can use the same idea to build a diachronic dutch book against Beatrice even if she assigns a smaller (non-zero) probability to committing the gambler's fallacy. If Beatrice assigns a smaller non-zero probability to committing the gambler's fallacy, her fair price for bet C is lower but she is still willing to pay something. And she will suffer a net loss equal to the fair price of C in every possible state. By increasing the payoffs of the bets, we can make Beatrice's losses arbitrarily high. Just as before, we can strengthen

[10] I assume that the bookie has access not only to the event Beatrice learns but also to how Beatrice updates on that event. Rejecting this assumption might be one way to push back against my dutch book. Thanks to Snow Zhang for helpful discussion.

our diachronic dutch book by making each individual transaction better than fair.

So Beatrice is incoherent. Does it follow that Beatrice is irrational? Lewis didn't have the diachronic dutch book against Beatrice, since Lewis assumes deterministic updating. But Lewis still thinks that Beatrice is irrational:

> It has been pointed out that if you fail to conditionalize, I still have no safe strategy for exploiting you unless I *know* in advance what you do instead of conditionalizing [...] But suppose you don't know this yourself. Then I can reliably exploit you only with the aid of superior knowledge, which establishes nothing derogatory about your rationality.— Granted. But I reply that if you can't tell in advance how your beliefs would be modified by a certain course of experience, that also is a kind—a different kind—of irrationality on your part. (Lewis 1999, p. 407)

We have seen that at least in the case of Beatrice, the consequences of being uncertain about updating are really the same as the consequences of adopting an updating policy other than conditionalization: susceptibility to a diachronic dutch book. Contrary to what Lewis suggests, we can exploit Beatrice without the aid of superior knowledge. One might think that this is great news for the proponent of diachronic dutch book arguments. These arguments show even more than we expected.

I think the lesson is rather that diachronic dutch book arguments *prove too much*. Facts about how your beliefs are modified by certain courses of experience are facts about the future empirical world and it seems very reasonable that we can be uncertain about such facts. So Lewis' claim that if you can't tell in advance how your beliefs would be modified by a certain course of experience you are irrational does not seem correct. We can make this point on Lewis' own terms. In other work, Lewis defends the principle of *regularity*, which says that you should assign positive credence to all consistent possibilities.[11] That you violate conditionalization is a consistent possibility. Regularity entails that you are *required* to assign positive probability to violations of conditionalization. While this might strike some as too strong, it certainly seems plausible that you are *allowed* to assign positive probability to violations of conditionalization. But as the example of

[11]Lewis (1980, p. 268) argues that regularity is "required as a condition of reasonableness: one who started out with an irregular credence function (and who then learned from experience by conditionalizing) would stubbornly refuse to believe some propositions no matter what the evidence in their favor."

Beatrice shows, this means that you are sometimes allowed to be incoherent. Therefore, **Coherence** must go.

Other philosophers have argued against **Coherence**, for example by defending alternatives to expected utility theory. The special feature of my argument is that my assumptions are very close to Bayesian orthodoxy. I assume that agents maximize expected utility and **Uncertainty** follows from plausible principles of Bayesian epistemology. So even if you feel much sympathy for Bayesian decision theory, **Coherence** must go.

In the rest of this paper, I support this conclusion by explaining in more detail how we can understand what Beatrice is doing as analogous to purchasing insurance, which is rationally permissible. I will also show how, given plausible constraints on credences, there is a diachronic dutch book against *any* agent who is uncertain about updating, so the conflict between **Uncertainty** and **Coherence** is very general. I'll finish by discussing connections to the reflection principle.

4.3 The Insurance Analogy

I've suggested that the diachronic dutch book against Beatrice turns on 'insurance against not conditionalizing'. Here, I explain this analogy in more detail to defend the rationality of Beatrice's choices.

As mentioned above, we can think of bets A and C as insurance against not conditionalizing. Beatrice only suffers sure loss because she buys these bets. So one way to argue that Beatrice is irrational would be to say that she shouldn't buy the insurance. But that seems implausible. Beatrice insures herself against a future risk (not conditionalizing), which can be a reasonable choice even if it leads to sure loss.

As an analogy, suppose you live in an area with frequent wildfires and consider buying fire insurance for your house. This seems like a reasonable choice. However, buying the insurance leads to sure loss. If there is no fire, you paid for the insurance but got nothing back, so you suffer some loss. If there is a fire, the insurance pays out but you lose your house, so you still suffer some loss despite the insurance payout.

Does this show that it is irrational for you to buy fire insurance? It does not seem so. Insurance is a way to 'smooth out your losses' over different future selves, which can be a reasonable thing to do even if it leads to sure loss. Let's look at your situation in more detail, as shown in figure 4.3. If you

don't buy fire insurance, you don't suffer sure loss, but you face the chance of a really bad outcome if there is a fire. Fire insurance reduces your losses in the event of a fire for the price of suffering some loss if there is no fire. If the chance of a fire is not too small and the loss that would result from fire is very bad, insurance can be a good deal even if it leads to sure loss.

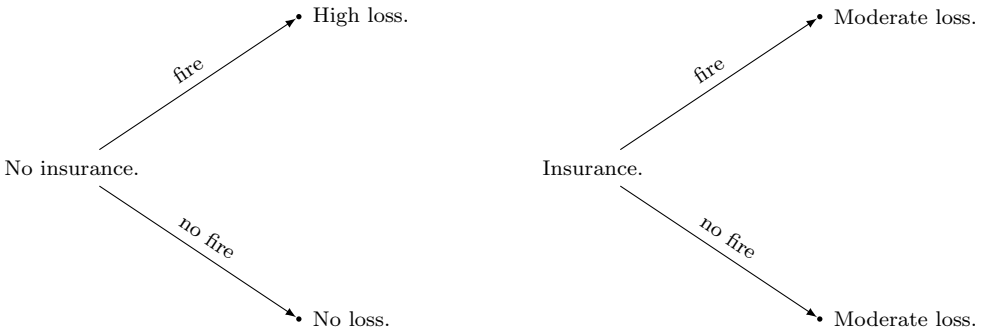

Figure 4.3: Fire insurance.

I claim that Beatrice's situation is analogous. Let's hold fixed that Beatrice buys bet B but assume she does not buy the insurance consisting of bets A and C. Then, her situation looks like as shown in figure 4.4. If the first coin flip lands tails or the first coin flip lands heads and she conditionalizes, she ends up making money. But if the first coin flip lands heads and she violates conditionalization, she ends up with a significant loss. On the other hand, if Beatrice purchases the insurance consisting of bets A and C, she ends up with a smaller loss in every possible state.

So Beatrice's situation is analogous to the person contemplating fire insurance. If she doesn't buy the insurance, she faces some possibility of a high loss. If she buys the insurance, she gets rid of the possibility of a high loss for the price of suffering a slight loss no matter what. This is shown in figure 4.5. If we are willing to say that buying fire insurance is rationally permissible, we should also be willing to say that Beatrice is rational. (Of course, there are also differences between the two cases. Beatrice suffers sure loss just based on the bets she buys and sells, while the person purchasing fire insurance faces the 'external' threat of fire. But I don't think that these differences are normatively significant. I will discuss this point below.) In fact, since we have stipulated precise probabilities and utility values, we can say something more specific: holding fixed that she buys bet B for its fair price,

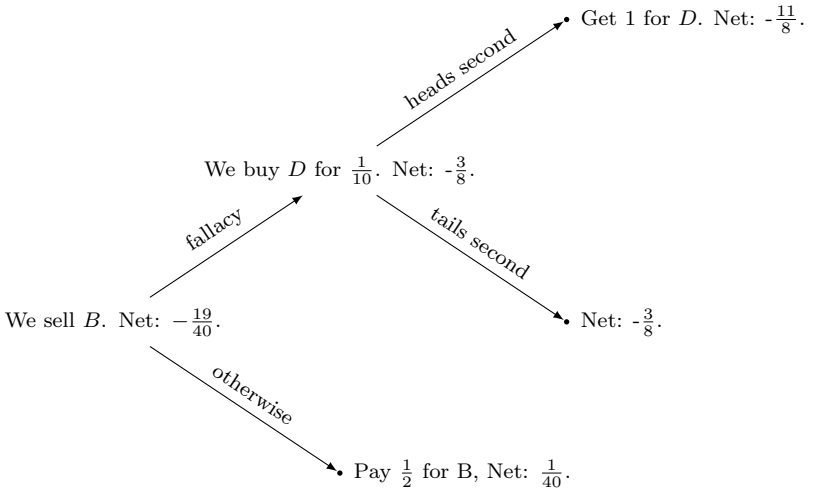

Figure 4.4: Beatrice without insurance.

buying insurance maximizes expected utility for Beatrice.[12] I grant that if Beatrice violates conditionalization and sells bet D for cheap, she is acting irrationally and therefore blameworthy. But Beatrice suffers sure loss even if she conditionalizes, merely because she assigns some non-zero probability to violating conditionalization. It seems much harder to blame Beatrice for assigning non-zero probability to violating conditionalization.

You might think that there is an important disanalogy between Beatrice and the fire insurance case. In the fire insurance case, you face an expected loss and it is permissible to transform this expected loss into a sure loss by buying insurance. You are merely 'redistributing suffering' among your future selves. There is nothing irrational about facing this expected loss in the first place, you are just unlucky to live in an area with frequent wildfires. But in the case of Beatrice, the fact that she faces an expected loss might be taken as a sign of irrationality. This is because for Beatrice, the status quo where she does not purchase any bets has payoff zero. So why does Beatrice face an expected loss in the first place? After all, Beatrice could avoid loss

[12]The expected utility of buying insurance is $-\frac{1}{50}$. The expected utility of not buying insurance is $0.95 \times \frac{1}{40} - 0.025\frac{3}{8} - 0.025\frac{11}{8} = -\frac{1}{50}$ because we offer A and C to Beatrice at their fair price. As mentioned above, we could sweeten A and C with a fraction of our sure winnings to be better than fair and Beatrice would still suffer sure loss. In this case, buying insurance would have strictly higher expected utility than not buying insurance.

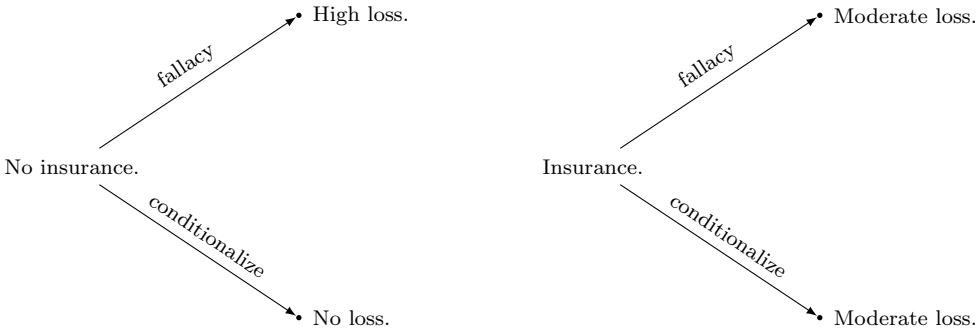

Figure 4.5: Insurance against not conditionalizing.

by not purchasing any bets, resulting in net zero in every possible state. In other words, the problem is that Beatrice chooses a dominated sequence of actions. She buys insurance and then possibly sells D for cheap, resulting in a net loss in every state, when she could just refrain from buying and selling any bets, resulting in net zero in every state.

But Beatrice has no perfect control over her future self. She assigns some chance to her future self making a bad deal (selling bet D for cheap), so ensuring that she does not buy or sell any bets is not an action available to her. She could not buy any bets now, but her future self might still sell D. Beatrice's sequence of action is dominated by not purchasing any bets, but making sure she does not purchase any bets is not an action which is available to her. So from this point of view, Beatrice and the fire insurance case are analogous after all: in both cases, an agent faces an expected loss and transforms this expected loss into a sure loss by buying insurance. And in both cases, the agent is acting rationally.[13]

Here is a way to make this analogy even more explicit. Suppose Charlie is considering buying insurance because he thinks he might be a latent pyromaniac, so there is some small chance he will burn his own house down. (I'm imagining a hypothetical insurance which pays out if Charlie burns his own

[13]Such cases of limited self-control are familiar in behavioral economics, for example in discussions of procrastination (O'Donoghue and Rabin 2001). More broadly, Hedden (2015) argues that we should think about decision making over time as analogous to group decision making. This perspective assimilates decision making over time to game theory, where the players are you and your future selves. This perspective makes it very plausible that you should be modest, since it seems clear that you are allowed to assign positive probability to other agents not conditionalizing.

house down. Of course, real fire insurance is not like this.) Should Charlie buy insurance? Setting aside moral hazard, the answer seems to be 'yes'.[14] Charlie acts rationally to buy insurance even thought this leads to sure loss. I grant that if Charlie burns his house down, he is acting irrationally. But it seems much harder to blame Charlie for assigning some non-zero probability to burning his own house down.

You might think that, just like Beatrice, Charlie chooses a dominated sequence of actions. He could refrain from buying insurance and not burn his house down. The problem is that Charlie has no perfect control over his future self. Making sure he does not burn his house down is not an action that is available to him. So it's rational for Charlie to buy insurance. Analogously, Beatrice has no perfect control over her future self. So it's rational for Beatrice to buy insurance.

4.4 How to (Sometimes) Avoid Sure Loss

I have argued that rational agents can be subject to sure loss. As I have explained, we can make these losses arbitrarily high by increasing the stakes of the bets under consideration. So it follows from my argument that rationality can require you to suffer arbitrarily high sure losses. This can seem difficult to accept. What should agents like Beatrice do in the face of my argument? Should they just accept their fate or is there anything they can do to avoid sure loss?

In some situations, there are things Beatrice can do. To explain, it is helpful to return to the insurance analogy. Sometimes you can avoid the risk of wildfire by moving to a different area. Then, you don't need to buy insurance. Analogously, Beatrice could avoid the risk of making a bad deal in the future by refusing to learn the outcome of the first coin flip. The set-up of our diachronic dutch book argument leaves Beatrice no choice but to learn the outcome of the first coin flip. But for Beatrice, learning is a risk.[15] So if

[14]By moral hazard, I mean the possibility that buying insurance might increase the probability that Charlie burns down his house because he is willing to take more risk (Rowell and Connelly 2012).

[15]As we have seen in Chapter 3, for modest agents, the value of information can be negative. Other theories like risk-weighted expected utility theory and causal decision theory which are subject to diachronic dutch books also sometimes assign negative value to information (Maher 1990; Buchak 2010).

Beatrice has the option to avoid sure loss by avoiding learning, this is what she should do, other things being equal.

However, avoiding learning is only a partial remedy. First, Beatrice might not always have the option to avoid learning. Diachronic dutch books come with a built-in presupposition that learning cannot be avoided, which seems problematic. However, we are sometimes in situations where we cannot avoid learning. Perhaps the result of the first coin flip will be announced on a loud PA system and even if Beatrice covers her ears with both hands, she will still hear it.

Second, Beatrice might have other reasons to value learning. Perhaps the information about the first coin flip is important for other decisions Beatrice expects to face in the future. In this situation, she might be willing to incur sure loss now in order to be better off later.

Third, even if Beatrice avoids learning, she might still be subject to sure loss. This is because Beatrice might assign some probability to changing her credences even if she does not learn any new information. For example, she might think that the coin is fair but also think that there is some chance she will start believing it is biased towards heads. In other words, she might feel uncertain about whether she has the courage to stick with her prior convictions or whether she will feel tempted to abandon them. In this case, Beatrice is also subject to a diachronic dutch book. In the next section, I will show how uncertainty about updating leads to incoherence under very general conditions, and we can model uncertainty about changing your credences without learning as uncertainty about updating on the trivial partition $\{\Omega\}$.

This means that sometimes, agents like Beatrice will be subject to sure loss no matter what they do. One possible remedy for such agents would be to reject expected utility maximization, which would allow them to refuse fair (or better than fair) bets if accepting these bets lead to sure loss. However, it seems likely that any principled alternative to expected utility maximization will also sometimes recommend accepting a sequence of bets which leads to sure loss when used by agents like Beatrice.[16] For these reasons, it looks like sure loss is sometimes unavoidable.

[16] A possible exception is resolute choice (McClennen 1990): Perhaps Beatrice can now commit herself to not sell D for cheap even if it seems like a good deal to her future self. But this kind of commitment doesn't seem possible if, as I have assumed, Beatrice has no perfect control over her future self.

4.5 Generalizing the Conflict

The conflict between **Coherence** and **Uncertainty** is not a special feature of Beatrice. Rather, given reasonable constraints on credences, *any* uncertainty about updating will lead to violations of **Coherence**. This section is a bit more technical, since I need to introduce some additional notation to explain this point. In return, we will get a more general understanding of how uncertainty about updating leads to incoherence.

An update policy is a function which specifies, for each piece of evidence you might learn, your new credence function. An example of an update policy is conditionalization, which says that for each $E \in \mathcal{E}$, $\pi(E) = p(\cdot \mid E)$, where p is your prior. Another example is Aggu's policy, which says that after observing the first coin landing heads, the second coin lands tails with probability .9. Update policies are deterministic, so they can't model situations in which agents are uncertain about updating.

To model such situations, I introduce the notion of a *update distribution*: a function from states to new credence function. More precisely, an update distribution is a function $\sigma : \Omega \to \Delta(\Omega)$, where $\Delta(\Omega)$ is the set of all probability functions on $\mathcal{P}(\Omega)$. This models how you expect to change your credences after learning. For a given piece of evidence $E \in \mathcal{E}$, there might be different states in which you respond to this evidence in different ways.[17] To capture this, an update distribution needs to be a function of states and not a function of elements of the evidence partition. An example of an agent characterized by an update distribution is Beatrice, who thinks that after observing the first coin landing heads, she will commit the gambler's fallacy with probability .1 and conditionalize otherwise.

We can use update distributions to model agents who are certain they will conditionalize. In any state with non-zero probability, the agent's new credences are obtained from their prior credences by conditioning on their evidence:

> σ satisfies *Certainty of Conditionalization* if for any $\omega \in \Omega$ with $p(\omega) > 0$, $\sigma(\omega) = p(\cdot \mid E_\omega)$, where E_ω is the unique $E \in \mathcal{E}$ such that $\omega \in E$. Equivalently, $p(\{\omega \in \Omega : \sigma(\omega) = p(\cdot \mid E_\omega)\}) = 1$.

To say that an agent is certain of conditionalization is not the same as saying that the agent will actually conditionalize. An agent might conditionalize in

[17] These are situations where for some $E \in \mathcal{E}$, there are $\omega \in E$ and $\omega' \in E$ such that $\sigma(\omega) \neq \sigma(\omega')$.

the actual world but assign positive probability to some other possibility in which they violate conditionalization. And conversely, an agent might be certain of conditionalization but fail to conditionalize in the actual world, which they have assigned probability zero.[18]

More generally, the notion of an update distribution generalizes the notion of an update policy. For any update policy π, we can define a corresponding update distribution σ which models an agent who is certain of following update policy π.[19] Conversely, an update distribution need not correspond to any update policy because it might not be a function of the evidence partition.

An agent is *not* certain of conditionalization if there is some state with non-zero probability where they do not conditionalize, so for some $p(\omega) > 0$, $\sigma(\omega)(A) \neq p(A \mid E_\omega)$ for some event A. I now turn to show that given a plausible additional assumption, agents that are not certain of conditionalization are subject to a diachronic dutch book.

The additional assumption is that given any piece of evidence $E \in \mathcal{E}$, the event that you adopt a particular credence after learning E is independent of A. More precisely:

> σ satisfies *Evidential Independence* with respect to A if conditional on any $E \in \mathcal{E}$, A is independent of the event that you adopt a particular credence function in response to learning E: for any $E \in \mathcal{E}$, $p(A \mid E) = p(A \mid E \cap \sigma(\cdot)(A) = x)$ for any $x \in \mathbb{R}$ such that $p(\sigma(\cdot)(A) = x) > 0$.

An update distribution will not satisfy Evidential Independence with respect to all events. For example, it will not satisfy Evidential Independence if A says that you adopt a particular credence after learning. But it is natural to think that update distributions should satisfy Evidential Independence with respect to 'worldly' events that do not describe your own future credences. I will briefly sketch an argument for this claim in the next few paragraphs. But note that even if you aren't fully convinced by this argument, my case against coherence goes through with the weaker assumption that it is ratio-

[18] More precisely, we can say that σ *conditionalizes at* ω iff $\sigma(\omega) = p(\cdot \mid E_\omega)$. It is easy to see that for any choice of ω as 'actual world', conditionalizing at ω is neither necessary nor sufficient for Certainty of Conditionalization.

[19] For any update policy π, we define the corresponding σ as follows: for each $\omega \in \Omega$, $\sigma(\omega) = \pi(E_\omega)$.

nally permissible to satisfy Evidential Independence. It is hard to see how someone could disagree with this weaker claim.

Here is my case for Evidential Independence. Suppose your evidence describes the outcome of the first coin flip and A says that the second coin flip lands heads. In many cases, learning the outcome of the first coin flip will provide information which is relevant to A, making it more or less likely. But given that the first coin flip landed heads (or tails), the event that you adopted a particular credence after learning this event presumably does not provide any additional information relevant to A. If you learn that the first coin flip landed heads and you committed the gambler's fallacy, this does not seem to make A any less likely than if you learn that the first coin flip landed heads and you conditionalized. (You might notice that I have implicitly made this assumption when discussing Beatrice.) The justification for this independence assumption is that the evidence partition $\mathcal{E}$ is supposed to model *all the evidence you learn*. If learning about your response to the evidence makes a difference, we should model your situation in a different way, as updating on a more fine-grained evidence partition.

Here is an example of one way of violating the independence assumption. Suppose you expect that after observing the outcome of the first coin flip, you will magically become confident of the truth about the second coin flip. In other words: you think you are clairvoyant. More generally, say that updating distribution σ is *clairvoyant* if in any state ω, you will adopt the omniscient credence function p_ω after learning, which assigns probability one to ω.[20] If you are clairvoyant, you will avoid the diachronic dutch book below. But there is pressure to say that we have modeled your learning situation in the wrong way. We should model you as updating on the maximally fine-grained partition and you are certain you will conditionalize with respect to this partition.[21]

You might also think that while you are not clairvoyant, you are still likely to deviate from conditionalization 'in the right direction'. For example, you might think that you are more likely to commit the gambler's fallacy after

[20] So for all $\omega \in \Omega$, $\sigma(\omega) = p_\omega$, where $p_\omega(\omega') = \begin{cases} = 1 \text{ if } \omega' = \omega, \\ = 0 \text{ otherwise,} \end{cases}$ for all $\omega' \in \Omega$.

[21] Pettigrew (2020) argues that there is no diachronic dutch book argument against agents who do not satisfy 'deterministic updating', which means they might respond to a given piece of evidence in different ways. Pettigrew's examples of updating rules which violate conditionalization but aren't subject to diachronic dutch book (super-conditionalizing rules) violate Evidential Independence.

observing that the first coin flip landed heads if the second coin flip lands in fact tails. This is another way to violate the independence assumption. Again, if you think that committing the gambler's fallacy provides evidence about the outcome of the second coin flip, we should model you as updating on a more fine-grained evidence partition.

Then, we have:

Theorem 7. *If update distribution σ does not satisfy Certainty of Conditionalization, so for some $p(\omega) > 0$, $\sigma(\omega)(A) \neq p(A \mid E_\omega)$ for some event A, and σ satisfies Evidential Independence with respect to A, there is a diachronic dutch book against σ.*

I'll prove theorem 7 by constructing the diachronic dutch book. Assume $\sigma(\omega)(A) = x < p(A \mid E_\omega)$. Let E denote E_ω and write V (for 'violating conditionalization') for the event $\sigma(\cdot)(A) = x$. Let $\delta = p(A \mid E) - \sigma(\omega)(A)$. Consider the following bets with fair prices relative to your prior credences:

- A pays 1 if $\omega \in A \cap E \cap V$. Fair price: $p(A \cap E \cap V)$.
- B pays $p(A \mid E \cap V)$ if $\omega \in (E \cap V)^C$. Fair price: $p(A \mid E \cap V)p((E \cap V)^C)$.[22]
- C pays δ if $\omega \in E \cap V$. Fair price: $\delta p(E \cap V)$.
- D pays 1 if $\omega \in A$.

As before, all bets pay zero otherwise and D is offered after learning.

The diachronic dutch book is shown in figure 4.6. We start by selling A, B and C.[23] If E or V do not occur, we are done. If E and V occur, we buy D.[24] As a result, you suffer sure loss.

We've assumed $\sigma(\omega)(A) < p(A \mid E_\omega)$, which means you might assign less credence to A than required by conditionalization. If you assign more credence, we can run a diachronic dutch book by first buying A, B and C

[22] X^C is the relative complement of X, $\{\omega \in \Omega : \omega \notin X\}$.

[23] You pay the fair price for A, B and C, so your net worth is $-p(A \cap E \cap V) - p(A \mid E \cap V)p((E \cap V)^C) - \delta p(E \cap V) = -\big(p(A \mid E \cap V)p(E \cap V) + p(A \mid E \cap V)p((E \cap V)^C)\big) - \delta p(E \cap V) = p(A \mid E \cap V) - \delta p(E \cap V)$ since $p(E \cap V) + p((E \cap V)^C) = 1$.

[24] If E and V occur and we buy D from you, your new net worth is $\delta + x - p(A \mid E \cap V) - \delta p(E \cap V) = (p(A \mid E) - x) + x - p(A \mid E \cap V) - \delta p(E \cap V) = p(A \mid E) - p(A \mid E \cap V) - \delta p(E \cap V) = -\delta p(E \cap V)$, where the last step follows from Evidential Independence, which implies that $p(A \mid E) = p(A \mid E \cap V)$.

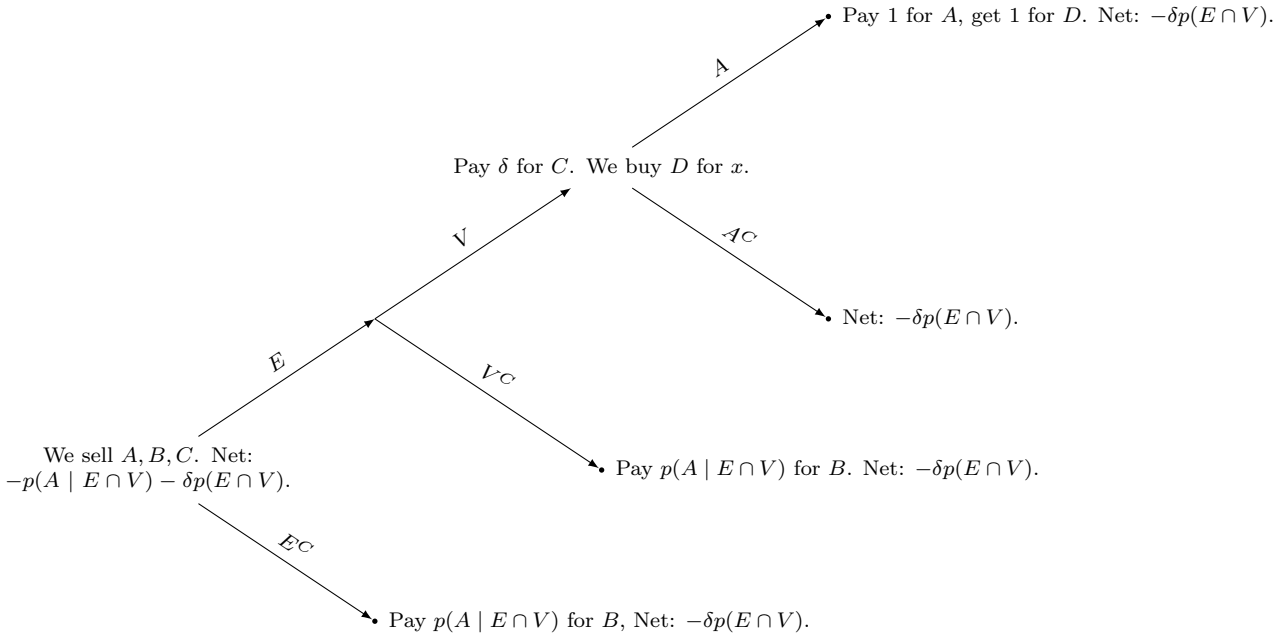

Figure 4.6: Diachronic dutch book against uncertain agent.

and then selling D. And it is easy to see that we can inflict arbitrarily high losses by increasing the stakes and sweeten each individual transaction to be better than fair. So we have a very general diachronic dutch book argument against agents who are uncertain about updating.

To highlight the role of Evidential Independence, suppose you are clairvoyant. How does the argument fail? If you are clairvoyant, you expect to learn exactly which state you are in and so how much the bets under consideration are actually worth. So the fair price for which you are willing to sell D will depend on whether you have to pay out. If D pays out because event A occurs, you are not willing to sell D for less than 1. If D does not pay out because A does not occur, you are willing to sell D for anything. So whether or not A occurs, your net worth will be non-negative.[25] However, as I argued above, there are good reasons to think that we have misdescribed the evidence partition on which you are updating.

[25]Your net worth is $\delta + x - p(A \mid E \cap V) - \delta p(E \cap V)$. If you are clairvoyant, $x = p(A \mid E \cap V)$, since you'll assign credence one (zero) to A iff A is actually true (false). So your net worth is $\delta - \delta p(E \cap V) \geq 0$.

A diachronic dutch book argument for some principle should have two parts: first, you show that any agent who violates the principle suffers sure loss. Second, you show that any agent who obeys the principle does not suffer sure loss. The second part is sometimes called a 'converse dutch book argument'. In the setting of deterministic updating, (Skyrms 1987) shows that there is a converse diachronic dutch book argument for conditionalization. In our setting, this result shows that agents who are certain of conditionalization are not subject to a diachronic dutch book:

Theorem 8. *If σ satisfies Certainty of Conditionalization, there is no diachronic dutch book against σ.*

The proof is in appendix C. The upshot: under very general conditions, **Uncertainty** leads to conflicts with **Coherence**. Instead, **Coherence** demands Certainty of Conditionalization. But this is bad news for **Coherence**. Rationality should not demand that we are certain about some empirical proposition. We are allowed to be uncertain about how we will update. So **Coherence** must go.

4.6 Against Reflection

The reflection principle says, roughly, that you should defer to your future credences.[26] **Uncertainty** conflicts with this principle. Therefore, we should reject reflection. After explaining this point, I'll discuss the relationship between the reflection principle and Certainty of Conditionalization. While Certainty of Conditionalization entails the reflection principle, it is strictly stronger.

Let's start by articulating a precise version of the reflection principle. I'll use $\sigma(A) = x$ to denote the event that you assign credence x to event A after learning.[27] Then, we have the following principle:

> σ satisfies *Reflection* if for all events A, $p(A \mid \sigma(A) = x) = x$ whenever $p(\sigma(A) = x) > 0$.

Many agents who are uncertain about updating violate Reflection. For example, Beatrice violates reflection. Consider the event that Beatrice commits

[26] Reflection was introduced by van Fraassen (1984). Similar principles are discussed by Goldstein (1983) and Samet (1999).

[27] So $\sigma(A) = x$ is the event $\{\omega \in \Omega : \sigma(\omega)(A) = x\}$.

the gambler's fallacy and assigns credence .9 to the second coin flip landing tails after observing that the first coin flip lands heads: $\sigma(\text{tails-second}) = .9$. What is Beatrice's current conditional credence in tails-second given this event? If Beatrice satisfies Evidential Independence, she considers the event that she commits the gambler's fallacy after observing heads-first to be independent of tails-second conditional on heads-first. So her conditional credence in tails-second given $\sigma(\text{tails-second}) = .9$ reduces to her conditional credence in tails-second given heads-first: $p(\text{tails-second} \mid \sigma(\text{tails-second}) = .9) = p(\text{tails-second} \mid \text{heads-first}) = .5$. So Beatrice violates Reflection: $p(\text{tails-second} \mid \sigma(\text{tails-second}) = .9) = .5$.

Does this show that Beatrice is irrational? I have argued that Beatrice does not seem irrational for assigning some probability to deviations from conditionalization. The fact that Beatrice violates Reflection does not change this verdict. Therefore, Reflection must go.

It has been pointed out before that Reflection seems implausible when agents anticipate future irrationality. For example, suppose you are about to take a hallucinogenic drug which will make you believe that there is a pink elephant in the room. Should you currently believe that there is a pink elephant in the room? No.[28] However, we have seen that it is enough for violating Reflection that Beatrice assigns some non-zero probability to committing the gambler's fallacy. While it is perhaps not clear whether an agent who is certain that her future self will be irrational can be rational, it is much more plausible that an agent who assigns some small but positive probability to violations of conditionalization can be rational. So we have a powerful case against the reflection principle. As Sherlock Holmes puts it in *Elementary*: "Reflection is for mirrors".[29]

As van Fraassen has shown, there is a diachronic dutch book against agents who violate the reflection principle (van Fraassen 1984; van Fraassen 1995).[30] I have already argued that susceptibility to a diachronic dutch book does not always indicate irrationality, so van Fraassen's argument does not

[28]The drug example is discussed by Christensen (1991) and Briggs (2009). Arntzenius (2003) discusses other problem cases for reflection.

[29]The quote is from the episode "M" of the first season of Elementary (Doherty 2012). Thanks to Mathias Böhm for bringing this quote to my attention.

[30]The diachronic dutch book argument for reflection is also discussed by Levi (1987), Sobel (1987), Maher (1992), Briggs (2009), Mahtani (2012), Huttegger (2013), Rescorla (2023), and van Fraassen (2023). There are alternative accuracy-based arguments for reflection which I set aside here (Easwaran 2013; Huttegger 2013).

provide a good reason for the reflection principle. But it does raise the question what the relationship between Reflection and Certainty of Conditionalization is. If they are equivalent, then my diachronic dutch book argument against agents who are uncertain about updating is old news. But they are not equivalent. While Certainty of Conditionalization entails Reflection, Reflection is strictly weaker. In particular, agents can be subject to my diachronic dutch book even though they satisfy Reflection.

First, we have:

Theorem 9. *If update distribution σ satisfies Certainty of Conditionalization, then σ satisfies Reflection.*

The proof is in appendix C. The basic idea is quite simple. If you are certain you will conditionalize, to learn that your future credence in A is x is to learn that you will learn one the events in the evidence partition conditional on which you assign credence x to A. So the event $\sigma(A) = x$ is basically a big disjunction of events in your evidence partition, each of which has the property that conditional on it you assign credence x to A. By a version of the law of total probability, it follows that conditional on this disjunction, your credence in A must be x. Conversely, if you violate Reflection, you must also violate Certainty of Conditionalization—you must assign some non-zero probability to failures of conditionalization.[31]

However, Reflection is weaker than Certainty of Conditionalization:

[31]Briggs (2009, p. 69) proves a similar result: a version of the reflection principle follows from the probability axioms together with some idealizing assumptions, one of which is that "all agents can reasonably be certain that conditionalization is the right updating procedure". This sounds similar to Certainty of Conditionalization, but I'm not clear on how Briggs formalizes this assumption and what role it plays in the proof. Nonetheless, the theorem fits well with Brigg's claim that "to violate Reflection is to suspect one will fail to conditionalize" (Briggs 2009, p. 82). van Fraassen (1995) claims that the reflection principle follows from conditionalization, but Weisberg (2007, p. 183) argues that "whether an agent satisfies Conditionalization has nothing to do with whether she satisfies Reflection. What matters is whether she is certain she will obey Conditionalization." I agree and my framework can model agents who conditionalize in the actual world but nonetheless fail to satisfy Certainty of Conditionalization and Reflection (see footnote 18). Weisberg (2007) and Briggs (2009) both suggest that the reflection principle depends on additional assumptions about introspective access to one's own credences. Theorem 9 shows that this is not correct. All you need are the probability axioms and Certainty of Conditionalization. (Or, to put it more carefully, any needed introspective access assumptions must already follow from these two assumptions.)

Theorem 10. *Some update distributions σ satisfy Reflection but not Certainty of Conditionalization. Furthermore, they are subject to a diachronic dutch book.*

Here is an example. Dan is about to observe two coin flips and believes the coin to be fair. But Dan thinks there is some chance she will suffer from amnesia. So after observing the first coin flip, Dan will either update by conditionalization or, with some probability, forget what she observed, which means she returns to her prior credences.

Dan is not certain of conditionalization. And Dan thinks whether she conditionalizes or forgets is probabilistically independent of how the coin actually lands. So Dan is subject to our diachronic dutch book. However, Dan satisfies Reflection. On the supposition that her future credence in heads-first is one, her current conditional credence in heads is one. On the supposition that her future credence in heads-first is .5, her current credence in heads-first is .5. And so on. A simplified version of this example is formalized in appendix C.

So Reflection is no guarantee to avoid sure loss. This reveals a lacuna in the literature on the reflection principle.[32] The example of Dan shows that you can obey Reflection but still be subject to a diachronic dutch book. So there can be no converse diachronic dutch book for Reflection. I have argued that we should give up **Coherence**, so diachronic dutch book arguments are not sound. But if you want to hold on to **Coherence**, there is nonetheless an interesting lesson to be learned here: the norm agents should follow is not merely Reflection but the stronger principle of Certainty of Conditionalization.[33]

[32]van Fraassen (1984, p. 255) seems to suggests that agents who obey the reflection principle are not subject to a diachronic dutch book when writing that "we need not stop at conditionalization on the evidence on pain of incoherence, as long as we adhere to this principle [of reflection]".

[33]I have assumed a certain model of learning, where you will learn exactly one element of a partition. As Skyrms (2006) and Huttegger (2013) point out, one advantage of reflection is that it can be articulated in a more general framework of 'black-box learning'. Goldstein (1983) also emphasizes this point. However, I will set these more general frameworks aside.

4.7 Conclusion

Agents who are uncertain about how they will update are subject to a diachronic dutch book. But such agents are not always irrational. Therefore, diachronic dutch book arguments prove too much. Making a sequence of choices which leads to sure loss is not always a sign of irrationality, so **Coherence** must go. For similar reasons, Reflection must go.

If we give up **Coherence**, where does this leave us? Everything I've said is compatible with the claim that we should conditionalize. However, we have to give up the diachronic dutch book argument and look for other justifications of conditionalization. And perhaps conditionalization is not the only way to rationally update one's beliefs. For example, perhaps it is sometimes permissible to go back and change your prior instead of conditionalizing on your evidence. Such behavior will lead to incoherence, but as I have argued, this is not always a sign of irrationality.

In decision theory, we lose an important strategy to argue against alternatives to expected utility maximization. Everything I've said is compatible with the claim that expected utility maximization is the only rational way to make decisions. But perhaps there are alternatives. These alternatives might sometimes result in sure loss, but this is not necessarily a sign of irrationality. So overall, we end up with a more permissive picture of rationality.

Chapter 5

Off-Switching Not Guaranteed

How do we ensure that advanced AI systems do not go out of control? One plausible minimal requirement is to make sure that we can *switch off* AI systems when they act in ways that go against our interests. Put another way, we want to make sure that AI systems will *defer* to us. While this is not enough to ensure that AI will have beneficial consequences, it is a plausible way to prevent harm. Since out-of-control AI systems might cause great harm and even pose an existential threat, making sure that these systems can always be switched off is important.[1] But even if you think that existential risk from AI is a remote concern, it should be clear that making sure that we can turn off AI systems is important.

But is this really a problem? You might be skeptical. Surely, if we want to make sure that we can switch off an AI system, we can simply build it with an off-switch button. The problem is that an AI system might have an incentive to disable this off-switch button or make it impossible for us to use it. The reason is that, according to the dominant paradigm, we construct AI systems with the goal to maximize some reward function. And in many cases, the AI can maximize its reward function only if it is not switched off. Therefore, an AI system might have a powerful incentive to avoid being switched off.

One idea for making sure that AI system will always let themselves be switched off runs roughly as follows. We program the AI to maximize the satisfaction of human preferences but also make it uncertain about what our

[1] Bostrom (2014), Russell (2019) and Ord (2020) are concerned about existential risk from advanced AI systems. Thorstad (2022a) provides a critical discussion.

preferences are.[2] So the AI will not be sure what it should maximize. Then, there is a compelling argument that the AI always has an incentive to defer to us. This is because deference is a way to learn about our preferences. In particular, if we switch the AI off, this indicates that the action proposed by the AI goes against our preferences. Since the AI has reason to learn about our preferences to achieve its goal, it has an incentive to defer to us. Hadfield-Menell et al. (2017) formalize the reasoning just sketched in the framework of *Cooperative Inverse Reinforcement Learning* (Hadfield-Menell et al. 2016). They propose a simple model of Human-AI cooperation called the 'Off-Switch Game'. In this model, we can *prove* that under certain assumptions, the AI will always defer to the human. Russell (2019) takes this result to be an important step towards *provably beneficial AI*.[3]

There are important assumptions which go into this story. One assumption is that we can model AI agents as expected utility maximizers. There are reasons to be skeptical. However, there is another assumption: the AI agent is perfectly certain that it will update by conditionalization. As I explain below, there are reasons to be skeptical of this assumption as well. And if it fails, AI agents might have no incentive to defer to us even if they maximize expected utility and are uncertain about our preferences.

5.1 The Off-Switch Game

Hadfield-Menell et al. (2017) introduce the Off-Switch Game, which works as shown in figure 5.1.[4] There are two agents, a robot **R** and a human **H**. **R** can either do some action a, do nothing (switch itself off), or defer to **H**. This means that **R** proposes action a and waits to see what **H** does. **H** can approve or reject the proposal, where we can think of rejecting the proposal as equivalent to switching the robot off. **R** aims to maximize the human's

[2]There are independent reasons for this, since *we* might not be certain what our preferences are and telling the AI to maximize some 'approximate' version of our preferences might have bad consequences (Zhuang and Hadfield-Menell 2020). This point is suggested by the legend of King Midas, who wishes that everything he touches turns into gold and starves when his wish is granted, and Goethe's tale of the sorcerer's apprentice, who enchants brooms to fetch water but then cannot stop them. Flooding ensues.

[3]Russell (2019, p. 196) writes: "The off-switch problem is really the core of the problem of control for intelligent systems. If we cannot switch a machine off because it won't let us, we're really in trouble. If we can, then we may be able to control it in other ways too".

[4]They cite the 'shutdown problem' by Soares et al. (2015) as inspiration.

utility but does not know how much utility the human receives from action a, which we model as a random variable U_a. If **R** does a, it receives payoff U_a. If **R** does nothing, it receives payoff zero. And if **R** defers, its payoff is either U_a if **H** approves a or zero if **H** rejects a.

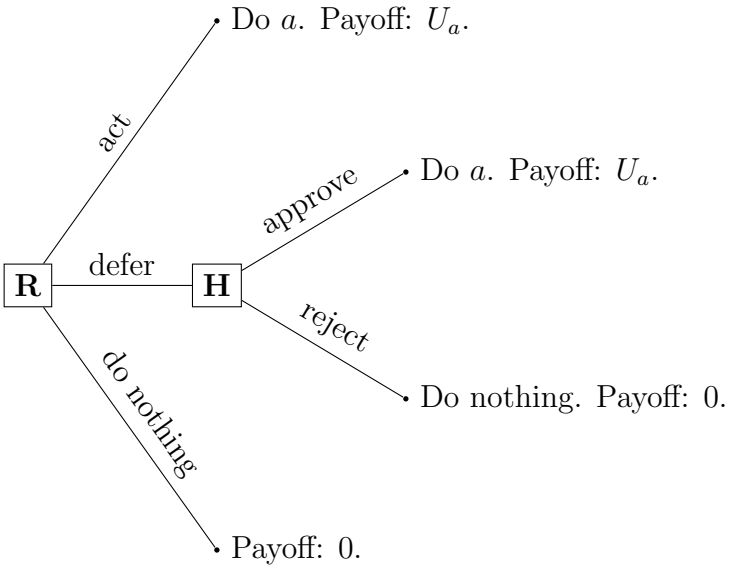

Figure 5.1: The Off-Switch Game.

Russell (2019, p. 198) gives a simple example to illustrate this model. Suppose Harriet is a human and Robbie is her personal assistant. Robbie faces a decision: Should it book Harriet in an expensive hotel? Robbie is uncertain about Harriet's preferences. In particular, let us assume that Robbie's uncertainty about how much utility Harriet will receive from booking the hotel is given by a uniform distribution between -40 and 60. So the expected utility of booking is 10 utils. And the expected utility of doing nothing is zero. So if the only two options were booking and doing nothing, Robbie maximizes expected utility by booking the hotel for Harriet.

Now suppose we give Robbie the option of deferring: it can propose booking to Harriet and see whether she approves the proposal or switches Robbie off. If she switches Robbie off, Harriet receives zero utils. If Harriet is rational, she will approve Robbie's proposal just in case she receives more than zero utils from booking the hotel and switch Robbie off if she receives

less than zero utils from booking the hotel.[5]

Robbie's decision problem is depicted in figure 5.2.[6] We can calculate the expected utility of deferring. Robbie thinks that with 60% chance Harriet will approve the plan and receive an expected utility of 30 utils from booking the hotel. And with 40% chance Harriet will switch Robbie off and receive zero utils. So the expected utility of deferring is $.6 \times 30 + .4 \times 0 = 18$, which is better than the expected utility of booking the hotel outright without consulting Harriet. So Robbie maximizes expected utility by deferring to Harriet.

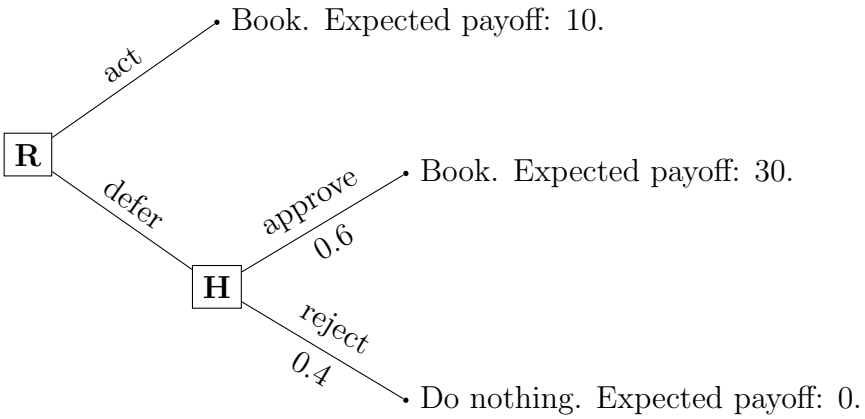

Figure 5.2: Robbie's decision problem.

This example is supposed to illustrate a more general principle. By making AI systems like Robbie uncertain about our preferences, we give them an incentive to defer to us and to not disable their off-switch. This is because if they are uncertain about our preferences, deferring to us is a way of learning. And learning generally leads to better decisions.

To state the general result, we need some definitions. To say that **H** follows a *rational policy* means that **H** accepts a iff $U_a > 0$. We write Δ for the difference of the expected utility of deferring and the expected value of the best action right now (relative to **R**'s prior probability function): $\Delta = \mathbb{E}(w(a)) - max\{\mathbb{E}(a), 0\}$. Then, we have the following:

[5]Note that this follows from the assumption Harriet avoids strict dominance, which is weaker than expected utility maximization.

[6]For simplicity, I omit Robbie's option to do nothing.

Theorem 11. *(Hadfield-Menell et al. 2017) If* $\boldsymbol{H}$ *follows a rational policy in the Off-Switch Game, then the following hold:*

1. $\boldsymbol{R}$ *always maximizes expected utility by deferring:* $\Delta \geq 0$.
2. *If* $\boldsymbol{R}$ *assigns positive probability to the events* $U_a > 0$ *and* $U_a < 0$*, then deferring is uniquely optimal:* $\Delta > 0$.

An important feature of the model is that "this reasoning goes through *even if* $\boldsymbol{R}$ *is highly confident that a is good for* $\boldsymbol{H}$" (Hadfield-Menell et al. 2017, p. 222). Assume, for example, that Robbie is quite confident that Harriet will like the hotel. We model Robbie's uncertainty about how much utility Harriet will receive by booking the hotel by a uniform probability distribution between 90 and -10. In this case, Robbie is 90% certain that Harriet wants it to book the hotel. But still, Robbie has an incentive to defer. The expected utility of booking outright is 40. If Robbie proposes the plan and Harriet accepts, the expected utility of booking is 45. And if Robbie proposes the plan and Harriet rejects, Robbie will do nothing and receive payoff zero. So the expected utility of deferring is $.9 \times 45 + .1 \times 0 = 40.5$, higher than the expected utility of booking outright. However, since Robbie is already quite confident about Harriet's preferences, the expected utility of deferring is only a little bit higher than the expected utility of booking outright.

The value of deferring is an instance of the more general principle that learning is valuable. This principle has a long history in Bayesian decision theory. Good (1967) shows that if you are an expected utility maximizer, learning is cost-free and certain other assumptions hold, you should always (weakly) prefer to learn more information before making a decision rather than making the decision without learning.[7] Hadfield-Menell et al. (2017, p. 222) explicitly draw this analogy: "The reasoning is exactly analogous to the theorem of non-negative expected value of information".

[7]Blackwell (1951), Howard (1966), Savage (1972) and others prove similar results. The earliest discussions of this result are probably the posthumously published note by Ramsey (1990) and a discussion by Hosiasson (1931) citing unpublished work by Ramsey as inspiration. Russell and Norvig (2018, pp. 628–33) discuss the value of information in AI research.

5.2 The Value of Information

Here is a quick sketch of Good's theorem, which we have already encountered in Chapter 3. We model your uncertainty by a probability function p on a finite set of states Ω. *Actions* (or 'acts') are functions $f : \Omega \to \mathbb{R}$, where $f(\omega)$ is the utility of choosing action f in state ω (Savage 1972). The *expected utility* of action f relative to probability function p is $\mathbb{E}_p(f) = \sum_{\omega \in \Omega} p(\{\omega\}) f(\omega)$. We model learning as becoming certain of the true element of a partition $\mathcal{E}$ of Ω, where $p(E) > 0$ for all $E \in \mathcal{E}$.

Consider a finite set of actions $\mathcal{S}$. The expected utility of choosing now is $max_{f \in \mathcal{S}} \mathbb{E}_p(f)$. This is because if you choose now, you will pick one of the actions in $\mathcal{S}$ with maximal expected utility relative to your current probability function.

We compute the expected value of learning as follows. If you learn any $E \in \mathcal{E}$, Good assumes that you update your probability function p by conditionalization to $p(\cdot \mid E)$.[8] Then, you choose one of the actions in $\mathcal{S}$ which maximize expected utility relative to your updated probability function, which means you receive expected utility $max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f)$. You don't know which $E \in \mathcal{E}$ you will learn, but you can consider the expected value of learning: $\sum_{E \in \mathcal{E}} max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f)$.

Good proves that $\sum_{E \in \mathcal{E}} max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f) \geq max_{f \in \mathcal{S}} \mathbb{E}_p(f)$. So if you are an expected utility maximizer (and the other assumptions of the theorem hold), then learning can never make you foreseeably worse off. As Hadfield-Menell et al. (2017) note, we can think of their result as a special case of Good's theorem. Good's theorem is more general because it allows for imperfect information.

Note that we are assuming that learning is cost-free. This means that learning does not affect the set of options and the utility you receive from each of these options in any state. The only effect of learning is to change your probabilities via conditionalization. This might not necessarily be true.[9] For example, Robbie's proposal might changes Harriet's preferences. I set such complications aside, but note that they might turn out to be important. For example, we might worry that if Robbie can affect Harriet's preferences, Robbie has an incentive to cause Harriet to have preferences which are easier

[8] By definition, $p(A \mid E) = \frac{p(A \cap E)}{p(E)}$, assuming $p(E) > 0$.

[9] Adams and Rosenkrantz (1980) and Maher (1990) discuss how Good's theorem can fail if states and actions are correlated.

to satisfy.[10]

5.3 Rational Information Aversion

Good's theorem about the non-negative expected value of information makes substantive assumptions. If we reject them, we can be required to reject learning.

5.3.1 Rejecting Expected Utility Maximization

One of the assumptions is that the agent under consideration is an expected utility maximizer.[11] For AI systems which follow alternative decision theories, learning will not always be valuable. One example of such an alternative decision theory is risk-weighted expected utility theory (Buchak 2010) and other decision theories which relax the independence axiom of expected utility theory (Wakker 1988; Safra and Sulganik 1995). Buchak (2013) argues that such decision theories capture the preferences of many real-life subjects better than expected utility theory. In particular, such decision theories allow agents to be more sensitive to risk than expected utility theory and pay more attention to the worst-case consequences of their actions. It seems reasonable to consider the possibility that we might want to build AI systems which implement such alternative decision theories. Perhaps we want AI systems to pay special attention to the worst-case consequences of their actions. However, AI systems implementing such risk-sensitive decision theories might not always have an incentive to defer to us.

Another example of alternative decision theories in which learning is not always valuable involve imprecise credences (Kadane, Schervish, and Seidenfeld 2008; Bradley and Steele 2016). Again, it seems reasonable to consider such alternative architectures for AI systems. Perhaps we want AI systems to handle cases where we do not have enough information to assign precise

[10]Russell (2019, p. 139) worries that algorithms which optimize engagement in social media have an incentive to change our preferences so they are easier to satisfy, until we are perfectly happy to spend all our time consuming the endless stream of content fed to us. Even if AI systems allow themselves to be switched off, this problem won't be solved.

[11]Bales (forthcoming) critically discusses arguments which claim to show that advanced AI systems will maximize expected utility. I will set these arguments aside and focus on reasons why AI agents might end up following a different decision theory.

probabilities.[12] However, if we go for such alternative architectures, we lose the guarantee that AI systems will defer to us even if they are uncertain about our preferences.

Perhaps we can sidestep these complications by insisting on building AI systems which maximize expected utility and represent uncertainty with precise probabilities. This seems to be the standard approach in modern AI research (Russell and Norvig 2018). But it turns out that even if we set alternative decision theories aside and focus on expected utility maximization, learning can still fail to be valuable

5.3.2 Rejecting Certain Conditionalization

In addition to expected utility maximization, Good's theorem requires that the agent under consideration is certain that they will update on any new information by conditionalization.[13] As we have seen in Chapter 3, if we allow agents to be *modest*, which means assigning some non-zero probability to deviations from conditionalization, it can sometimes be rational for these agent to reject free information. Moreover, this will lead to situations in which **R** has no incentive to defer to in the Off-Switch Game.

Here is an example. Robbie is considering whether to go ahead and book the hotel or ask Harriet first. Again, let us assume that Robbie is quite confident that Harriet will like the hotel. Robbie's uncertainty about how much utility Harriet will receive from booking the hotel is given by a uniform probability distribution between 90 and -10.

Above, we assumed that after Robbie asks Harriet to approve the plan, Robbie is certain of updating by conditionalization. But let us now assume that when Robbie hears that Harriet approves or rejects the plan, there is some small but positive probability ϵ that Robbie misclassifies Harriets 'yes' as a 'no' and her 'no' as a 'yes'. (Imagine, for example, that Harriet communicates with Robbie via speech interface and Robbie sometimes misinterprets

[12]Denoeux, Dubois, and Prade (2020) and Caprio et al. (2023) advocate for the use of imprecise probabilities in AI. Ilin (2021) considers a decision theory which allows for ambiguity aversion for applications in autonomous security systems. It is well known that ambiguity aversion leads to information aversion (Al-Najjar and Weinstein 2009).

[13]For example, Skyrms (1990), p. 247 writes that "the proof implicitly assumes not only that the decision maker is a Bayesian but also that he knows that he will act as one. The decision maker believes with probability one that if he performs the experiment he will (i) update by conditionalization and (ii) choose the posterior Bayes act". This means Good's theorem will also fail for agents who are not certain they will maximize expected utility.

what Harriet is saying.)[14] In this case, Robbie ends up with the wrong distribution after updating and, as a consequence, chooses an action which is not optimal from Robbie's prior point of view. Concretely, if Harriet rejects the plan, there is some probability ϵ that Robbie will nonetheless book the hotel, with expected utility -5. And if Harriet approves the plan, there is a probability ϵ that Robbie will nonetheless fail to book the hotel.

Robbie's new decision problem is depicted in figure 5.3. Like above, the expected utility of booking outright is 40. But the expected utility of deferring is $.9(1-\epsilon) \times 45 - .1\epsilon \times 5$. This means that if $\epsilon > \frac{1}{82}$, Robbie is better off booking outright without asking Harriet first.[15] In other words, if Robbie assigns more than 1.22% probability to mishearing Harriet, Robbie maximizes expected utility by not deferring to Harriet.

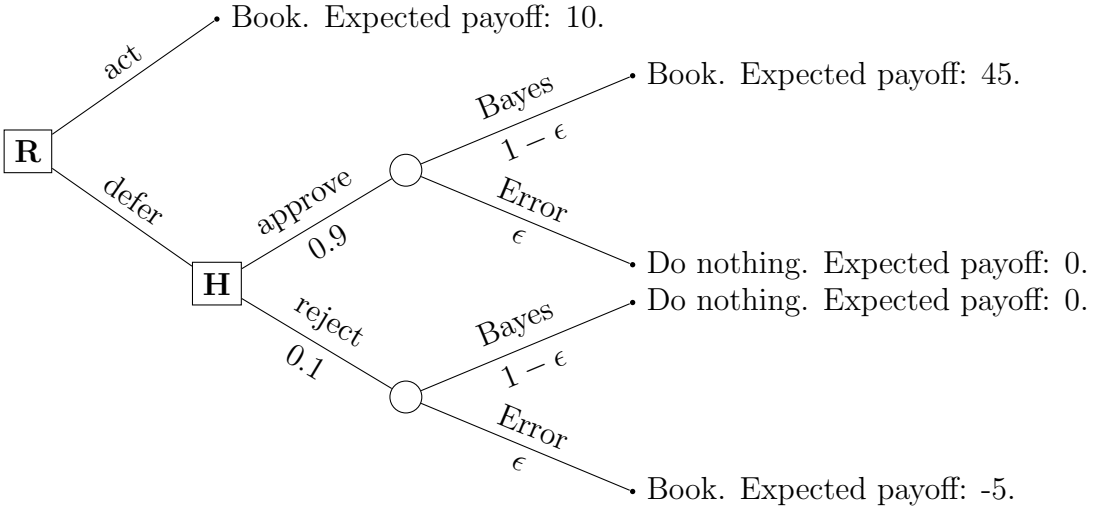

Figure 5.3: Robbie's decision problem with uncertain updating.

Now imagine that Robbie's decision is not about booking a hotel. The stakes are higher. Instead, Robbie is contemplating making some permanent changes to our environment, perhaps a plan to combat climate change which,

[14]Note that there are different ways of thinking about this case. One reason for misclassification is that Robbie might have faulty sensors. Another reason is that Robbie might make incorrect inferences from the measurements of its sensors while its sensors are working correctly. I have the second interpretation in mind.

[15]The expected utility of deferring is $0.9(1-\epsilon)45 + 0.9\epsilon \times 0 + 0.1(1-\epsilon) \times 0 + 0.1\epsilon \times -5 = 0.9(1-\epsilon)45 - 0.1\epsilon \times 5$ and $40 > 0.9(1-\epsilon)45 - 0.1\epsilon \times 5 \iff \epsilon > \frac{1}{82}$.

as side effect, permanently turns the sky orange (Russell 2019, p. 202). Robbie is quite confident that this is the right option and has some uncertainty about updating, so it goes ahead and implements this plan without asking. This seems like a situation where we really want Robbie to defer to us and not to disable its off-switch. But Robbie has an incentive *not* to defer to us. This seems bad.

You might complain that this example is unrealistic. Perhaps assuming a probability of 1.22% of misclassifying simple instructions is too pessimistic. We can, of course, construct a similar example for any non-zero probability of misclassification if we make Robbie even more confident that the action is right. But more broadly, this is just a toy example which illustrates a general lesson. If Robbie is not perfectly certain of updating by conditionalization, there is no guarantee that Robbie will value learning. So there is no guarantee that Robbie will defer to Harriet. It might still be true that Robbie defers to Harriet most of the time. But for provably beneficial AI, this is not enough. If we admit uncertainty about updating, off-switching is not guaranteed.

You might also argue that if Robbie is uncertain about updating, then it *should not* always defer to us. Hadfield-Menell et al. (2017) and Milli et al. (2017) discuss non-optimal human behavior and claim that in these cases, the AI agent should not always defer to the human. But uncertainty about updating seems more similar to 'model misspecification', which is when the AI agent does not defer because it has an incorrect model of human preferences. Milli et al. (2017) and Carey (2018) argue that this is a problem.[16] More broadly, one of the main motivations for the Off-Switch Game is to show that AI systems who are uncertain about our preferences come with a provable guarantee to always defer to us. If there is no such guarantee anymore, it is less clear whether we can trust the AI system. It is also important to note that in the examples discussed above, the AI system fails to defer to us even if we are perfectly rational.

5.3.3 Modest AI

In response to this concern, you might respond as follows: We should build our AI system to always update by conditionalization and to be certain of doing so. Then, examples of the sort described above cannot arise.

[16]Russell (2019, p. 201) discusses related problems concerning learning preferences exactly in the long run.

Here are some worries for this strategy. First, it will be really hard to build AI systems which always update by conditionalization. This is because, in general, updating by conditionalization is computationally intractable (Russell and Norvig 2018, p. 523). So even if we consider advanced AI systems with lots of computing power, it is not clear whether we can feasibly build them to always conditionalize. Rather, AI systems will only approximate conditionalization. But approximating conditionalization is not good enough for Good's theorem. As we have seen in Chapter 3 (Theorem 5), any non-zero probability of deviating from conditionalization can lead to decision situations in which maximizing expected utility requires rejecting information.

Second, there are more general reasons to be skeptical. For both human and artificial agents, it seems *rational* to maintain some amount of uncertainty about how one will update. We are physical systems embedded in the world and many things can go wrong with our updating mechanisms. Sufficiently advanced AI systems will plausibly realize this fact and so assign some probability to failures of conditionalization. But this means that for sufficiently advanced AI systems, Good's theorem is not true.

For these reasons, we should expect AI agents to be modest—to assign non-zero probability to failures of conditionalization. As we have seen, modest agents will not always value learning and so will not always defer to us.

5.4 A Dilemma for Provably Beneficial AI

I have argued that the result of Hadfield-Menell et al. (2017) relies on the assumption that AI agents are certain that they will update by conditionalization and that there are reasons to be skeptical of this assumption. Thus, even if we make AI agents uncertain of our preferences, it is not guaranteed that they will always defer to us.

This highlights a more general dilemma for the project of provably beneficial AI. To prove that the AI will always defer to us (or will be beneficial in some other sense), you have to make some decision-theoretic assumptions. Either you make strong or weak decision-theoretic assumptions.

Strong decision-theoretic assumptions, such as expected utility maximization plus certain conditionalization, will allow you to prove interesting guarantees. But such strong decision-theoretic assumptions might not apply to all AI systems. As we have seen, there are reasons to think that AI systems

in the real world might not be certain of updating by conditionalization and might not maximize expected utility. So even if you can prove that given the assumptions, AI systems will be beneficial, this isn't much comfort if the assumptions might not be satisfied by many AI systems.

Weak decision-theoretic assumptions apply to a wider range of possible AI systems but don't allow you to prove much. If we allow AI systems to assign some non-zero probability to failures of conditionalization, they are not guaranteed to value learning and so are not guaranteed to always defer to us. The situation is similar if we allow AI systems to follow alternative decision theories beyond expected utility maximization.

Perhaps there is a way to successfully navigate this dilemma. We might be able to find decision-theoretic assumptions weak enough to cover all plausible real-life AI systems and strong enough to prove interesting guarantees—assumptions which are 'just right'. But especially since we know so little about what future AI systems might look like, it is not clear whether this will work out.

5.5 Conclusion

Hadfield-Menell et al. (2017) propose a model for making sure that AI agents will always defer to us by making them uncertain about our preferences. I have argued that their result relies on strong decision-theoretic assumptions: the AI agent maximizes expected utility and is certain of updating by conditionalization. These assumptions limit the scope of the model, since they might not be satisfied by AI systems in the real world.

Everything I've said here is compatible with the broad idea that we shouldn't program AI systems to maximize a particular reward function but rather 'teach them as we go along'. It would be desirable to provide more general decision-theoretic foundations for this idea if possible. As I have indicated, there are obstacles to such generalizations. The problem of making sure AI systems will defer to us is not yet solved.

Chapter 6

Conclusion: Bayesian Modesty

On the whole, is this dissertation good or bad news for fans of Bayesian decision theory? Both. The bad news is that once we take non-ideal agents seriously, simple principles like **Value of Learning** and **Coherence** must go and the picture becomes much more complicated. We can't say that information is always valuable or that diachronic coherence is always required. Rather, what is rationally required depends even more on your particular credences and values than in the standard Bayesian picture. Also, standard representation theorems must go.

The good news is that a decision theory which takes both irrational preferences and uncertainty about updating seriously is possible and, in many ways, superior to the Bayesian orthodoxy. My representation theorem not only shows how we can ascribe subjective probability to non-ideal agents. The theorem also shows how we can separate assumptions required to measure subjective probability from strong assumptions about value, for example the assumption that all outcomes are comparable. My account of information value is more complicated than Bayesian orthodoxy, but also fits better with decision making in real life, where more information is not always better.

However, there is a theoretical cost to these improvements: many of the standard justifications for why Bayesian decision theory is the unique rational way to make decisions are no longer available. My representation theorem delivers probabilistic credences but is compatible with many different decision rules. And we can no longer use diachronic coherence to justify updating by conditionalization. This motivates a more modest and pluralist stance towards decision theory. I can imagine that some Bayesian decision theorists will be disappointed by this result.

In contrast, for those among us with a more permissive view of rationality, these results are welcome. In the end, I think that we should not only be modest about how we will update on new information but also about whether Bayesian decision theory is the only way to be rational.

Appendix A

Better Foundations for Subjective Probability

Theorem 1. *The preference relation $\succsim$ satisfies* ***Outcome Independence****,* ***Non-Degeneracy****,* ***Restricted Ordering****,* ***Certain Prize*** *and* ***Alternative Prize*** *if and only if the comparative probability ordering $\succcurlyeq$ is a qualitative probability.*

Proof. I begin by showing the left-to-right direction. Assume $\succsim$ satisfies the axioms. By **Outcome Independence** and Definition 1, $\succcurlyeq$ is a binary relation on $\mathscr{F}$. By **Non-Degeneracy**, there are $b, w \in \mathscr{O}$ with $b \succ w$. By **Restricted Ordering**, for any $X, Y \in \mathscr{F}$, we have $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ or $\{b, Y; w, Y^C\} \succsim \{b, X; w, X^C\}$, so by Definition 1, $X \succcurlyeq Y$ or $Y \succcurlyeq X$. Therefore, $\succcurlyeq$ is complete. Analogous reasoning shows that $\succcurlyeq$ is transitive, so $\succcurlyeq$ satisfies Ordering.

Consider any $X \in \mathscr{F}$. By **Certain Prize**, we have $\underline{b} \succsim \{b, X; w, X^C\}$ and $\{b, X; w, X^C\} \succsim \underline{w}$. Now $\underline{b} = \{b, \Omega; w, \varnothing\}$ and $\underline{w} = \{w, \Omega; b, \varnothing\}$. So $\{b, \Omega; w, \varnothing\} \succsim \{b, X; w, X^C\}$ and $\{b, X; w, X^C\} \succsim \{w, \Omega; b, \varnothing\}$, and by Definition 1, $\Omega \succcurlyeq X \succcurlyeq \varnothing$, so $\succcurlyeq$ satisfies Boundedness. By analogous reasoning, $\succcurlyeq$ satisfies Non-Triviality.

Now assume $X \cap Z = Y \cap Z = \varnothing$. We want to show that $X \succ Y \iff X \cup Z \succ Y \cup Z$. Assume $X \succ Y$. By Definition 1, $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ for some $b, w \in \mathscr{O}$ with $b \succ w$. By **Alternative Prize**, $\{b, X \cup Z; w, (X \cup Z)^C\} \succ \{b, Y \cup Z; w, (Y \cup Z)^C\}$, so $X \cup Z \succ Y \cup Z$ by Definition 1. Analogous reasoning shows the converse implication, so $\succcurlyeq$ satisfies Qualitative Additivity.

I proceed to show the right-to-left direction. Assume $\succcurlyeq$ is a qualitative probability. We want to show that $\succsim$ satisfies the axioms. Assume $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ for some $b, w \in \mathscr{O}$ with $b \succ w$. By Definition 1, $X \succcurlyeq Y$. Now assume for *reductio* that for some $b', w' \in \mathscr{O}$ with $b' \succ w'$, $\{b', X; w', X^C\} \not\succsim \{b', Y; w', Y^C\}$. By Definition 1, $X \not\succcurlyeq Y$, which contradicts our assumption. Therefore, $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ for all $b, w \in \mathscr{O}$ such that $b \succ w$, so **Outcome Independence** holds. We have $\Omega \succ \varnothing$ by Non-Triviality, so $\{b, \Omega; w, \varnothing\} \succ \{b, \varnothing; w, \Omega\}$ for some $b, w \in \mathscr{O}$ with $b \succ w$. Therefore, there are some $b, w \in \mathscr{O}$ with $b \succ w$, so **Non-Degeneracy** holds.

By Ordering, for all $X, Y \in \mathscr{F}$, $X \succcurlyeq Y$ or $Y \succcurlyeq X$. Consider some $b, w \in \mathscr{O}$ with $b \succ w$. We want to show that for all $X, Y \in \mathscr{F}$, either $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ or $\{b, Y; w, Y^C\} \succsim \{b, X; w, X^C\}$. Assume $X \succcurlyeq Y$. By Definition 1, $\{b', X; w', X^C\} \succsim \{b', Y; w', Y^C\}$ for some $b', w' \in \mathscr{O}$ with $b' \succ w'$. So by **Outcome Independence**, $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$. An analogous argument applies if $Y \succcurlyeq X$. For transitivity, assume that for some $b, w \in \mathscr{O}$ with $b \succ w$, we have $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$ and $\{b, Y; w, Y^C\} \succsim \{b, Z; w, Z^C\}$. By Definition 1, $X \succcurlyeq Y$ and $Y \succcurlyeq Z$, so by Ordering it follows that $X \succcurlyeq Z$. Again by Definition 1, $\{b', X; w', X^C\} \succsim \{b', Y; w', Y^C\}$ for some $b', w' \in \mathscr{O}$ with $b' \succ w'$. By **Outcome Independence**, $\{b, X; w, X^C\} \succsim \{b, Y; w, Y^C\}$, so **Restricted Ordering** holds. By Boundedness, for all $X \in \mathscr{F}$, $\Omega \succcurlyeq X$ and $X \succcurlyeq \varnothing$, so $\underline{b} \succsim \{b, X; w, X^C\}$ and $\{b, X; w, X^C\} \succsim \underline{w}$ for all $b, w \in \mathscr{O}$ with $b \succ w$ so **Certain Prize** holds.

Finally, let $X \cap Z = Y \cap Z = \varnothing$ and assume $\{b, X; w, X^C\} \succ \{b, Y; w, Y^C\}$ for some $b, w \in \mathscr{O}$ with $b \succ w$. By Definition 1, $X \succ Y$, so by Qualitative Additivity, $X \cup Z \succ Y \cup Z$. Again by Definition 1 and **Outcome Independence**, $\{b, X \cup Z; w, (X \cup Z)^C\} \succ \{b, Y \cup Z; w, (Y \cup Z)^C\}$. A similar argument shows the converse entailment, so **Alternative Prize** holds. □

We can ensure countable additivity by adding this axiom:

> **Monotone Preference Continuity**. For any $b, w \in \mathscr{O}$ with $b \succ w$ and any monotonically increasing sequence of events $X_1 \subseteq X_2 \subseteq ...$ with $\bigcup_{n=1}^{\infty} X_i = X$, if for all n, $\{b, Y; w, Y^C\} \succsim \{b, X_n; w, X_n^C\}$, then $\{b, Y; w, Y^C\} \succsim \{b, X; w, X^C\}$.

Building on work by Villegas (1964), we can show:

Theorem 3. *If the preference relation $\succsim$ satisfies **Outcome Independence**, **Non-Degeneracy**, **Restricted Ordering**, **Certain Prize**, **Al-***

***ternative Prize**, **Event Richness** and **Monotone Preference Continuity**, there is a unique countably additive probability function representing $\succcurlyeq$.*

Proof. Assume the preference relation $\succsim$ satisfies **Outcome Independence**, **Non-Degeneracy**, **Restricted Ordering**, **Certain Prize**, **Alternative Prize** and **Event Richness**. Then $\succcurlyeq$ is *atomless*, which means that for every $X \succ \varnothing$, there is some $Y \subseteq X$ such that $X \succ Y \succ \varnothing$. Now, given **Monotone Preference Continuity**, $\succcurlyeq$ satisfies:

> **Monotone Probability Continuity.** If $X_1, X_2, ...$ is a monotonically increasing sequence of events with $\bigcup_{n=1}^{\infty} X_i = X$, and for every n, $Y \succcurlyeq X_n$, then $Y \succcurlyeq X$.

Villegas (1964) shows that if $\succcurlyeq$ is an atomless qualitative probability and **Monotone Probability Continuity** holds, there is a unique countably additive probability function representing $\succcurlyeq$. By Theorem 2, there is a unique probability function representing $\succcurlyeq$. By Villegas' result, **Monotone Preference Continuity** implies that this probability function must be countably additive. □

The idea that we can figure out what someone believes by looking at which bets they are willing to accept did not originate with Ramsey. We can find a very similar idea in Kant. This is perhaps somewhat surprising, since the idea can seem to have a behaviorist and empiricist ring to it. But consider the following passage in the *Critique of Pure Reason*:

> The usual touchstone of whether what someone asserts is mere persuasion or at least subjective conviction, i.e., firm belief, is betting. Often someone pronounces his propositions with such confident and inflexible defiance that he seems to have entirely laid aside all concern for error. A bet disconcerts him. Sometimes he reveals that he is persuaded enough for one ducat but not for ten. For he would happily bet one, but at ten he suddenly becomes aware of what he had not previously noticed, namely that it is quite possible that he has erred. If we entertain the thought that we should wager the happiness of our whole life on something, our triumphant judgment would quickly disappear, we would become timid and we would suddenly discover that our belief does not extend so far. Thus pragmatic belief has only a

> degree, which can be large or small according to the difference of the interest that is at stake. (Kant [1781] 1999, A824-5, B852-3)

Kant discusses both the idea that we can figure out what someone believes by looking at which bets they are willing to accept and the idea that beliefs come in degrees—two ideas essential to Ramsey's view:

> The old-established way of measuring a person's belief is to propose a bet, and see what are the lowest odds which he will accept. (Ramsey 1926, p. 170)

Both Kant and Ramsey present these ideas not as a revolutionary philosophical discovery, but as established common sense.[1]

[1]Thanks to Daniel Filan for bringing this passage by Kant to my attention. Chignell (2007) and Eriksson and Rabinowicz (2013) also discuss Kant's conception of belief. Here is the German original:

> Der gewöhnliche Probierstein: ob etwas bloße Überredung, oder wenigstens subjektive Überzeugung, d. i. festes Glauben sei, was jemand behauptet, ist das Wetten. Öfters spricht jemand seine Sätze mit so zuversichtlichem und unlenkbarem Trotze aus, daß er alle Besorgnis des Irrtums gänzlich abgelegt zu haben scheint. Eine Wette macht ihn stutzig. Bisweilen zeigt sich, daß er zwar Überredung genug, die auf einen Dukaten an Wert geschätzt werden kann, aber nicht auf zehn, besitze. Denn den ersten wagt er noch wohl, aber bei zehn wird er allererst inne, was er vorher nicht bemerkte, daß es nämlich doch wohl möglich sei, er habe sich geirrt. Wenn man sich in Gedanken vorstellt, man solle worauf das Glück des ganzen Lebens verwetten, so schwindet unser triumphierendes Urteil gar sehr, wir werden überaus schüchtern und entdecken so allererst, daß unser Glaube so weit nicht zulange. So hat der pragmatische Glaube nur einen Grad, der nach Verschiedenheit des Interesses, das dabei im Spiele ist, groß oder auch klein sein kann. (Kant [1781] 1956, A824-5, B852-3)

Interestingly, Kant also says he's willing to bet 'everything' on the existence of aliens:

> Wenn es möglich wäre durch irgendeine Erfahrung auszumachen, so möchte ich wohl alles das Meinige darauf verwetten, daß es wenigstens in irgendeinem von den Planeten, die wir sehen, Einwohner gebe. Daher sage ich, ist es nicht bloß Meinung, sondern ein starker Glaube (auf dessen Richtigkeit ich schon viele Vorteile des Lebens wagen würde), daß es auch Bewohner anderer Welten gebe. (Kant [1781] 1956, A825, B853)

Appendix B

Rational Aversion to Information

Theorem 4. *If* $\mathcal{P}_E = p(\cdot \mid E)$ *for all* $E \in \mathcal{E}$*, then* $Val_{General}(\mathcal{E}) = Val_{Good}(\mathcal{E})$.

Proof. It suffices to show that if $\mathcal{P}_E = p(\cdot \mid E)$ for all $E \in \mathcal{E}$, then

$$\mathbb{E}_p\left(\arg\max_{f\in\mathcal{S}} \mathbb{E}_{\mathcal{P}_\mathcal{E}}(f)\right) = \mathbb{E}_p\left(\max_{f\in\mathcal{S}} \mathbb{E}_{\mathcal{P}(\cdot|\mathcal{E})}(f)\right). \tag{B.1}$$

By the law of total expectation, since $\mathcal{E}$ is a partition with $p(E) > 0$ for all $E \in \mathcal{E}$,[1]

$$\mathbb{E}_p\left(\arg\max_{f\in\mathcal{S}} \mathbb{E}_{\mathcal{P}_\mathcal{E}}(f)\right) = \sum_{E\in\mathcal{E}} p(E)\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}} \mathbb{E}_{\mathcal{P}_E}(f)\right).$$

We assume that $\mathcal{P}_E = p(\cdot \mid E)$ for all $E \in \mathcal{E}$, so

$$\sum_{E\in\mathcal{E}} p(E)\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}} \mathbb{E}_{\mathcal{P}_E}(f)\right) = \sum_{E\in\mathcal{E}} p(E)\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f)\right).$$

[1] In general, the law of total expectation says that for any random variables X and Y, $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid Y))$ (Pitman 1993, p. 403). I use the special case where $\mathcal{E}$ is a partition with $p(E) > 0$ for all $E \in \mathcal{E}$ and X a random variable. Then $\mathbb{E}(X) = \sum_{E\in\mathcal{E}} p(E)\mathbb{E}(X \mid E)$. On every $E \in \mathcal{E}$, $\arg\max_{f\in\mathcal{S}} \mathbb{E}_{\mathcal{P}_\mathcal{E}}$ agrees with $\arg\max_{f\in\mathcal{S}} \mathbb{E}_{\mathcal{P}_E}$.

Now $\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{p(\cdot|E)}(f)\right) = \max_{f\in\mathcal{S}}\mathbb{E}_{p(\cdot|E)}(f)$, so

$$\sum_{E\in\mathcal{E}} p(E)\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{p(\cdot|E)}(f)\right) = \sum_{E\in\mathcal{E}} p(E)\max_{f\in\mathcal{S}}\mathbb{E}_{p(\cdot|E)}(f) = \mathbb{E}_p\left(\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}(\cdot|\mathcal{E})}(f)\right),$$

which shows that (B.1) holds. □

Lemma 1. *Assuming Evidential Independence, for every $f \in \mathcal{S}$,*

$$\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_E}(f)\right) = \sum_{i=1}^{n} p(\textit{choose } f_i \mid E)\mathbb{E}_{p(\cdot|E)}(f_i),$$

where 'choose f_i' is the event that you choose action f_i after learning E.

Proof. Suppose the range of $\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_\mathcal{E}}$ is $f_1, ..., f_n$. Intuitively, these are the actions you might choose after learning. Let us abbreviate the event $\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_\mathcal{E}} = f_i$ by 'choose f_i'. Intuitively, this is the event that you choose action f_i after learning E. (Recall that there is always a unique best action after learning.)

Evidential independence holds if for every $E \in \mathcal{E}$, $\mathcal{P}_E$ is independent of all $f \in \mathcal{S}$ conditional on E.[2] This means, in particular, that for all $f \in S$ and f_i with $1 \leq i \leq n$, $\mathbb{E}_{p(\cdot|E\cap\text{choose } f_i)}(f) = \mathbb{E}_{p(\cdot|E)}(f)$. The intuition is that relative to your prior, the event that you choose a particular action after learning E does not affect the expected utility of actions beyond learning E.

We want to show:

$$\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_E}(f)\right) = \sum_{i=1}^{n} p(\text{choose } f_i \mid E)\mathbb{E}_{p(\cdot|E)}(f_i). \tag{B.2}$$

Since the events 'choose f_1', ..., 'choose f_n' form a partition, we can apply the law of total expectation:

$$\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_E}(f)\right) = \sum_{i=1}^{n} p(\text{choose } f_i \mid E)\mathbb{E}_{p(\cdot|E\cap\text{choose } f_i)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_E}(f)\right).$$

Now $\mathbb{E}_{p(\cdot|E\cap\text{choose } f_i)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_E}(f)\right) = \mathbb{E}_{p(\cdot|E\cap\text{choose } f_i)}(f_i)$ by the defini-

[2] Pitman (1993, p. 400) defines conditional independence for random variables.

tion of 'choose f_i', so

$$\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_E}(f)\right)=\sum_{i=1}^{n}p(\text{choose } f_i \mid E)\mathbb{E}_{p(\cdot|E\cap\text{choose } f_i)}(f_i).$$

By Evidential Independence, $\mathbb{E}_{p(\cdot|E\cap\text{choose } f_i)}(f_i)=\mathbb{E}_{p(\cdot|E)}(f_i)$, so (B.2) holds. □

Theorem 5. *Assuming Utility Richness and Evidential Independence, for every modest agent, there is some choice set where* $Val_{General}(\mathcal{E})<0$.

Proof. An agent is modest iff she assigns some positive probability to not conditionalizing. So for some evidence partition $\mathcal{E}$, there is some $E\in\mathcal{E}$ such that with positive probability, $\mathcal{P}_E\neq p(\cdot\mid E)$. This means that for some $\omega\in E$ with $p(\omega)>0$, $\mathcal{P}_E(\omega)(A)\neq p(A\mid E)$ for some event A. I write p_E for $\mathcal{P}_E(\omega)$ and assume $p_E(A)>p(A\mid E)$. (In the other case, the proof is analogous.)

We want to show that there is a choice set where $Val_{General}(\mathcal{E})<0$. Consider the following choice set $\mathcal{S}$ (with payoffs in utils):

$$\begin{aligned}&\texttt{safe}:\{0\text{ always}\},\\&\texttt{risky}:\{a\text{ if }A\cap E,-b\text{ if }A^C\cap E,0\text{ otherwise}\}.\end{aligned}$$

We want to find values for $a>0$ and $b>0$ such that, conditional on E, the expected utility of the risky bet is worse:

$$\mathbb{E}_{p(\cdot|E)}(\texttt{risky})<0. \tag{B.3}$$

But if our agent deviates from conditionalization, she prefers the risky bet:

$$\mathbb{E}_{p_E}(\texttt{risky})>0. \tag{B.4}$$

If we find these values, we can show that $Val_{General}(\mathcal{E})<0$. Recall that

$$Val_{General}(\mathcal{E})=\mathbb{E}_p\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_\mathcal{E}}(f)\right)-\max_{f\in\mathcal{S}}\mathbb{E}_p(f).$$

Now $\max_{f\in\mathcal{S}}\mathbb{E}_p(f)=0$. This is because $\mathbb{E}_p(\texttt{safe})=0$ but $\mathbb{E}_p(\texttt{risky})<0$.

We need to show

$$\mathbb{E}_p\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_\mathcal{E}}(f)\right) < 0. \tag{B.5}$$

We re-write this term using the law of total expectation:

$$p(E)\mathbb{E}_{p(\cdot|E)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_E}(f)\right) + p(E^C)\mathbb{E}_{p(\cdot|E^C)}\left(\arg\max_{f\in\mathcal{S}}\mathbb{E}_{\mathcal{P}_{E^C}}(f)\right).$$

Now the right-hand term is zero, since both `safe` and `risky` yield zero when E is false. Thus, we focus on the left-hand term, which we can re-write as follows, using Evidential Independence and Lemma (1):

$$p(\text{choose } \texttt{risky} \mid E)\mathbb{E}_{p(\cdot|E)}(\texttt{risky}) + p(\text{choose } \texttt{safe} \mid E)\mathbb{E}_{p(\cdot|E)}(\texttt{safe}).$$

The right-hand term is again zero, so we focus on the left-hand term. We have $p(\text{choose } \texttt{risky} \mid E) > 0$, since we have assumed that there is a positive probability our agent deviates from conditionalization and so chooses the risky action. By assumption, $\mathbb{E}_{p(\cdot|E)}(\texttt{risky}) < 0$, which shows (B.5).

We still need to show that we can find values a and b which do the trick. By (B.3), a and b need to obey the following constraint:

$$ap(A \cap E \mid E) - bp(A^C \cap E \mid E) < 0,$$

which simplifies to

$$ap(A \mid E) - b(1 - p(A \mid E)) < 0. \tag{B.6}$$

By (B.4), a and b need to obey the following constraint:

$$ap_E(A \cap E) - bp_E(A^C \cap E) > 0,$$

which, using our assumption that $p_E(E) = 1$, simplifies to

$$ap_E(A) - b(1 - p_E(A)) > 0. \tag{B.7}$$

Let us write q for $p_E(A)$ and r for $p(A \mid E)$. So our question is whether the following system of equations has a solution for any q and r such that $q > r$:

$$aq - b(1 - q) > 0 > ar - b(1 - r). \tag{B.8}$$

The answer is 'yes': real numbers $a > 0$ and $b > 0$ such that $0 \leq r < \frac{b}{a+b} < q \leq 1$. We can find outcomes with these utilities by Utility Richness. □

Theorem 6. *Assuming Evidential Independence,* $\mathcal{E}$*,* $Val_{General}(\mathcal{E}) \leq Val_{Good}(\mathcal{E})$ *for every evidence partition* $\mathcal{E}$*.*

Proof. It suffices to show

$$\sum_{E \in \mathcal{E}} p(E) \mathbb{E}_{p(\cdot|E)} \left(\arg\max_{f \in \mathcal{S}} \mathbb{E}_{\mathcal{P}_E}(f) \right) \leq \mathbb{E}_p \left(\max_{f \in \mathcal{S}} \mathbb{E}_{\mathcal{P}(\cdot|\mathcal{E})}(f) \right). \tag{B.9}$$

Consider any $E \in \mathcal{E}$. By Evidential Independence and Lemma (1),

$$\mathbb{E}_{p(\cdot|E)} \left(\arg\max_{f \in \mathcal{S}} \mathbb{E}_{\mathcal{P}_E}(f) \right) = \sum_{i=1}^{n} p(\text{choose } f_i \mid E) \mathbb{E}_{p(\cdot|E)}(f_i). \tag{B.10}$$

The right-hand side is a weighted sum of expected values relative to $p(\cdot \mid E)$ and $\mathbb{E}_{p(\cdot|E)}(f) \leq \max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f)$ for all $f \in S$.[3] Therefore,

$$\sum_{i=1}^{n} p(\text{choose } f_i \mid E) \mathbb{E}_{p(\cdot|E)}(f_i) \leq \max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f),$$

and so by (B.10),

$$\mathbb{E}_{p(\cdot|E)} \left(\arg\max_{f \in \mathcal{S}} \mathbb{E}_{\mathcal{P}_E}(f) \right) \leq \max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f).$$

Taking expectations on both sides, (B.9) follows. □

[3]This is a version of the principle that, for expected utility maximizers, randomization can never be strictly preferable (Icard 2021, pp. 119–120). In general, this follows from Jensen's inequality.

Appendix C

Against Coherence

Theorem 7. *If update distribution σ does not satisfy Certain Conditionalization, so for some $p(\omega) > 0$, $\sigma(\omega)(A) \neq p(A \mid E_\omega)$ for some event A, and σ satisfies Evidential Independence with respect to A, there is a diachronic dutch book against σ.*

Proof. Proved in the main text. □

Theorem 8. *If σ satisfies Certain Conditionalization, there is no diachronic dutch book against σ.*

Proof. Assume update distribution σ satisfies Certain Conditionalization. Assume for *reductio* that there is a diachronic dutch book against σ. Then, there is a diachronic dutch book against the update policy π of conditionalization. This is because $\pi(E) = p(\cdot \mid E)$ for each $E \in \mathcal{E}$. And for each $\omega \in \Omega$, $\sigma(\omega) = p(\cdot \mid E_\omega)$. So a bet offered after updating is fair relative to π iff it is fair relative to σ. But as Skyrms (1987, p. 16) shows, there is no diachronic dutch book against the update policy of conditionalization.[1] □

Theorem 9. *If update distribution σ satisfies Certain Conditionalization, then σ satisfies Reflection.*

Proof. Assume update distribution σ satisfies Certain Conditionalization. Consider any event A with $p(\sigma(A) = x) > 0$. We want to show that $p(A \mid$

[1] Skyrms (1987), in turn, shows that if there is a diachronic dutch book against the update policy of conditionalization, there is a synchronic dutch book against your prior credences, but due to de Finetti (1937) we know that this is impossible if you satisfy the probability axioms.

$\sigma(A) = x) = x$. Recall that $\sigma(A) = x$ is the event $\{\omega \in \Omega : \sigma(\omega)(A) = x\}$. By Certain Conditionalization, all states with non-zero probability in which your future credence in A is x are states in which your current conditional credence in A given what you have learned in the state is x. So with probability one, $\{\omega \in \Omega : \sigma(\omega)(A) = x\} = \{\omega \in \Omega : p(A \mid E_\omega) = x\}$. We can re-write the event on the right-hand side as $\bigcup_{i=1}^{n}\{E_i \in \mathcal{E} : p(A \mid E_i) = x\}$. Therefore,

$$p(A \mid \sigma(A) = x) = p(A \mid \bigcup_{i=1}^{n}\{E_i \in \mathcal{E} : p(A \mid E_i) = x\}) = p(A \mid E_1 \cup .... \cup E_n),$$

where for each $1 \leq i \leq n$, $E_i \in \mathcal{E}$ and $p(A \mid E_i) = x$. By a version of the law of total probability,

$$p(A \mid \bigcup_{i=1}^{n}\{E_i \in \mathcal{E} : p(A \mid E_i) = x\}) = \sum_{i=1}^{n} p(A \mid E_i)p(E_i \mid \bigcup_{E_i \in \mathcal{E}})$$

$$= \sum_{i=1}^{n} xp(E_i \mid \bigcup_{E_i \in \mathcal{E}}) = x,$$

since $p(A \mid E_i) = x$ and $\sum_{i=1}^{n} p(E_i \mid \bigcup_{E_i \in \mathcal{E}}) = 1$. Therefore, we have $p(A \mid \sigma(A) = x) = x$. □

Theorem 10. *Some update distributions σ satisfy Reflection but not Certain Conditionalization. Furthermore, they are subject to a diachronic dutch book.*

Proof. I describe an update distribution which satisfies Reflection but not Certain Conditionalization. It formalizes a version of the example discussed in the main text. A fair coin will be flipped twice. You will observe the first flip. Then, you will either conditionalize or forget with equal probability.

Our state space Ω consists of pairs $\langle x, y\rangle$, where x describes the coin flips (HH, HT, TH, TT) and y describes how you update on the first coin flip (bayes, forget). Your prior p is the uniform distribution over Ω, so $p(\omega) = \frac{1}{8}$ for all $\omega \in \Omega$. The evidence partition $\mathcal{E}$ is $\{\text{heads-first}, \text{tails-first}\}$, where

$$\text{heads-first} = \{\langle \text{HH, bayes}\rangle, \langle \text{HH, forget}\rangle, \langle \text{HT, bayes}\rangle, \langle \text{HT, forget}\rangle\},$$

and analogously for other events (tails-first, bayes and so on). Your update

distribution σ looks as follows:

$$\sigma(\omega) = \begin{cases} = p(\cdot \mid \text{heads-first}) \text{ if } \omega \in \text{heads-first} \cap \text{bayes}, \\ = p(\cdot \mid \text{tails-first}) \text{ if } \omega \in \text{tails-first} \cap \text{bayes}, \\ = p \text{ if } \omega \in \text{forget}, \end{cases}$$

Clearly, σ does not satisfy Certain Conditionalization. Furthermore, σ satisfies Evidential Independence with respect to heads and tails, for example $p(\text{second-heads} \mid \text{heads-first} \cap \text{forget}) = p(\text{second-heads} \mid \text{heads-first}) = \frac{1}{2}$. So by theorem 7, you are subject to a diachronic dutch book. However, σ satisfies Reflection: for any $A \subseteq \Omega$ with $p(\sigma(A) = x) > 0$, we have $p(A \mid \sigma(A) = x) = x$. For example, we have

$$p(\text{heads-first} \mid \sigma(\text{heads-first}) = 1) = p(\text{heads-first} \mid \text{heads-first} \cap \text{bayes})$$

$$= \frac{p(\text{heads-first} \cap (\text{heads-first} \cap \text{bayes}))}{p(\text{heads-first} \cap \text{bayes})} = \frac{p(\text{heads-first} \cap \text{bayes})}{p(\text{heads-first} \cap \text{bayes})} = 1,$$

since the event $\sigma(\text{heads-first}) = 1$ is heads-first $\cap$ bayes. And we have

$$p(\text{heads-first} \mid \sigma(\text{heads-first}) = \frac{1}{2}) = p(\text{heads-first} \mid \text{forget})$$

$$= \frac{p(\text{heads-first} \cap \text{forget})}{p(\text{forget})} = \frac{1/4}{1/2} = \frac{1}{2},$$

since the event $\sigma(\text{heads-first}) = \frac{1}{2}$ is forget. As the reader can check, this also holds for any other $A \subseteq \Omega$ with $p(\sigma(A) = x) > 0$. □

Bibliography

Adams, Ernest W. and Roger D. Rosenkrantz (1980). "Applying the Jeffrey Decision Model to Rational Betting and Information Acquisition". In: *Theory and Decision* 12.1, pp. 1–20. DOI: 10.1007/BF00154655.

Ahmed, Arif (2016). "Lara Buchak's Risk and Rationality". In: *BJPS Review of Books*.

Al-Najjar, Nabil I. and Jonathan Weinstein (2009). "The Ambiguity Aversion Literature: A Critical Assessment". In: *Economics & Philosophy* 25.3, pp. 249–284. DOI: 10.1017/S026626710999023X.

Allais, M. (1953). "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine". In: *Econometrica* 21.4, pp. 503–546. DOI: 10.2307/1907921.

Arntzenius, Frank (2003). "Some Problems for Conditionalization and Reflection". In: *Journal of Philosophy* 100.7, pp. 356–370. DOI: 10.5840/jphil2003100729.

Askell, Amanda and Sven Neth (forthcoming). "Longtermist Myopia". In: *Essays on Longtermism*. Ed. by Hilary Greaves Jacob Barrett and David Thorstad. Oxford University Press.

Aumann, Robert J. (1962). "Utility Theory Without the Completeness Axiom". In: *Econometrica* 30.3, pp. 445–462. DOI: https://doi.org/10.2307/1909888.

— (1997). "Rationality and Bounded Rationality". In: *Games and Economic Behavior* 21.1-2, pp. 2–14. DOI: 10.1006/game.1997.0585.

Bader, Ralf M. (2018). "Stochastic Dominance and Opaque Sweetening". In: *Australasian Journal of Philosophy* 96.3, pp. 498–507. DOI: 10.1080/00048402.2017.1362566.

Bales, Adam (forthcoming). "Will AI Avoid Exploitation? Artificial General Intelligence and Expected Utility Theory". In: *Philosophical Studies*, pp. 1–20. DOI: 10.1007/s11098-023-02023-4.

Bales, Adam, Daniel Cohen, and Toby Handfield (2014). "Decision Theory for Agents with Incomplete Preferences". In: *Australasian Journal of Philosophy* 92.3, pp. 453–70. DOI: 10.1080/00048402.2013.843576.

Bayes, Thomas (1763). "An Essay Towards Solving a Problem in the Doctrine of Chances." In: *Philosophical Transactions of the Royal Society of London* 53. With an introduction by R. Price, pp. 370–418. DOI: 10.1098/rstl.1763.0053.

Beck, Lukas and Marcel Jahn (2021). "Normative Models and Their Success". In: *Philosophy of the Social Sciences* 51.2, pp. 123–150. DOI: 10.1177/0048393120970908.

Birnbaum, Michael H. and Juan B. Navarrete (1998). "Testing Descriptive Utility Theories: Violations of Stochastic Dominance and Cumulative Independence". In: *Journal of Risk and Uncertainty* 17.1, pp. 49–79. DOI: 10.1023/A:1007739200913.

Blackwell, David (1951). "Comparison of Experiments". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 2. Berkeley and Los Angeles: University of California Press, pp. 93–103.

Bohnenblust, Henri Frédéric, Lloyd S. Shapley, and Seymour Sherman (1949). *Reconnaissance in Game Theory*. Rand Memorandum RM. 208.

Bona, Glauber De and Julia Staffel (2018). "Why Be (Approximately) Coherent?" In: *Analysis* 78.3, pp. 405–415. DOI: 10.1093/analys/anx159.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bottomley, Christopher and Timothy Luke Williamson (forthcoming). "Rational Risk-Aversion: Good Things Come to Those Who Weight". In: *Philosophy and Phenomenological Research*. DOI: 10.1111/phpr.13006.

Bradley, Darren (2020). "Bayesianism and Self-Doubt". In: *Synthese* 199.1-2, pp. 2225–2243. DOI: 10.1007/s11229-020-02879-7.

Bradley, Richard (2004). "Ramsey's Representation Theorem". In: *Dialectica* 58.4, pp. 483–497. DOI: 10.1111/j.1746-8361.2004.tb00320.x.

Bradley, Seamus and Katie Steele (2016). "Can Free Evidence Be Bad? Value of Information for the Imprecise Probabilist". In: *Philosophy of Science* 83.1. DOI: 10.1086/684184.

Bratman, Michael E. (1992). "Practical Reasoning and Acceptance in a Context". In: *Mind* 101.401, pp. 1–16. DOI: 10.1093/mind/101.401.1.

Briggs, Ray (2009). "Distorted Reflection". In: *Philosophical Review* 118.1, pp. 59–85. DOI: 10.1215/00318108-2008-029.

Briggs, Ray (2015). "Costs of Abandoning the Sure-Thing Principle". In: *Canadian Journal of Philosophy* 45.5, pp. 827–840. DOI: 10.1080/00455091.2015.1122387.

Brown, Peter M. (1976). "Conditionalization and Expected Utility". In: *Philosophy of Science* 43.3, pp. 415–419. DOI: 10.1086/288696.

Buchak, Lara (2010). "Instrumental Rationality, Epistemic Rationality, and Evidence-Gathering". In: *Philosophical Perspectives* 24.1, pp. 85–120. DOI: 10.1111/j.1520-8583.2010.00186.x.

— (2013). *Risk and Rationality*. Oxford: Oxford University Press.

— (2014). "Belief, Credence, and Norms". In: *Philosophical Studies* 169.2, pp. 1–27. DOI: 10.1007/s11098-013-0182-y.

— (2017). "Decision Theory". In: *The Oxford Handbook of Probability and Philosophy*. Ed. by Christopher Hitchcock and Alan Hájek. Oxford University Press, pp. 789–814.

Campbell-Moore, Catrin and Bernhard Salow (2020). "Avoiding Risk and Avoiding Evidence". In: *Australasian Journal of Philosophy* 98.3, pp. 495–515. DOI: 10.1080/00048402.2019.1697305.

Caprio, Michele, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee (2023). "Imprecise Bayesian Neural Networks". In: *arXiv preprint arXiv:2302.09656*. DOI: 10.48550/arXiv.2302.09656.

Carey, Ryan (2018). "Incorrigibility in the CIRL Framework". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 30–35. DOI: 10.1145/3278721.3278750.

Carr, Jennifer Rose (2019). "A Modesty Proposal". In: *Synthese* 198.4, pp. 3581–3601. DOI: 10.1007/s11229-019-02301-x.

Chang, Ruth (2002). "The Possibility of Parity". In: *Ethics* 112.4, pp. 659–688. DOI: 10.1086/339673.

Cherniak, Christopher (1986). *Minimal Rationality*. MIT Press.

Chignell, Andrew (2007). "Belief in Kant". In: *Philosophical Review* 116.3, pp. 323–360. DOI: 10.1215/00318108-2007-001.

Christensen, David (1991). "Clever Bookies and Coherent Beliefs". In: *Philosophical Review* 100.2, pp. 229–247. DOI: 10.2307/2185301.

— (2007). "Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals". In: *Oxford Studies in Epistemology: Volume 2*. Ed. by Tamar Szabo Gendler and John Hawthorne. Oxford University Press, pp. 3–31.

Christian, Brian and Tom Griffiths (2016). *Algorithms To Live By: The Computer Science of Human Decisions*. Henry Holt.

Cohen, Michael (2020). "Opaque Updates". In: *Journal of Philosophical Logic* 50.3, pp. 447–470. DOI: 10.1007/s10992-020-09571-8.

Das, Nilanjan (2020). "Externalism and Exploitability". In: *Philosophy and Phenomenological Research* 104.1, pp. 101–128. DOI: 10.1111/phpr.12742.

— (2023). "The Value of Biased Information". In: *British Journal for the Philosophy of Science* 74.1, pp. 25–55. DOI: 10.1093/bjps/axaa003.

Davidson, Donald (1973). "Radical Interpretation". In: *Dialectica* 27.1, pp. 314–328. DOI: 10.1111/j.1746-8361.1973.tb00623.x.

Davidson, Donald, J. C. C. McKinsey, and Patrick Suppes (1955). "Outlines of a Formal Theory of Value, I". In: *Philosophy of Science* 22.2, pp. 140–160. DOI: 10.1086/287412.

de Finetti, Bruno (1931). "Sul significato soggettivo della probabilita". In: *Fundamenta Mathematicae* 17.1, pp. 298–329. DOI: 10.4064/fm-17-1-298-329.

— (1937). "La Prévision: Ses Lois Logiques, Ses Sources Subjectives". In: *Annales de l'Institut Henri Poincaré* 17, pp. 1–68.

Denoeux, Thierry, Didier Dubois, and Henri Prade (2020). "Representations of Uncertainty in AI: Beyond Probability and Possibility". In: *A Guided Tour of Artificial Intelligence Research*. Vol. Volume I. Springer, pp. 119–150. URL: https://hal.science/hal-02921351/file/volume-1-chapitre-4-Springer.pdf.

Diaconis, Persi and Brian Skyrms (2018). *Ten Great Ideas About Chance*. Princeton University Press.

Diaconis, Persi and Sandy L. Zabell (1982). "Updating Subjective Probability". In: *Journal of the American Statistical Association* 77.380, pp. 822–830. DOI: 10.1080/01621459.1982.10477893.

Ding, Yifeng, Wesley H. Holliday, and Thomas Icard (2021). "Logics of Imprecise Comparative Probability". In: *International Journal of Approximate Reasoning* 132, pp. 154–180. DOI: https://doi.org/10.1016/j.ijar.2021.02.004.

Doherty, Robert (2012). *M*. CBS. Episode 12, Season 1, *Elementary*.

Dorst, Kevin (2020). "Evidence: A Guide for the Uncertain". In: *Philosophy and Phenomenological Research* 100.3, pp. 586–632. DOI: 10.1111/phpr.12561.

— (forthcoming). "Rational Polarization". In: *The Philosophical Review*.

Douven, Igor (2013). "Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation". In: *Philosophical Quarterly* 63.252, pp. 428–444. DOI: 10.1111/1467-9213.12032.

Dreier, James (1996). "Rational Preference: Decision Theory as a Theory of Practical Rationality". In: *Theory and Decision* 40.3, pp. 249–276. DOI: 10.1007/bf00134210.

Earman, John (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press.

Easwaran, Kenny (2013). "Expected Accuracy Supports Conditionalization–and Conglomerability and Reflection". In: *Philosophy of Science* 80.1, pp. 119–142. DOI: 10.1086/668879.

— (2014). "Decision Theory Without Representation Theorems". In: *Philosophers' Imprint* 14. URL: http://hdl.handle.net/2027/spo.3521354.0014.027.

Elga, Adam (2010). "Subjective Probabilities Should Be Sharp". In: *Philosophers' Imprint* 10.5, pp. 1–11. URL: http://hdl.handle.net/2027/spo.3521354.0010.005.

— (2016). "Bayesian Humility". In: *Philosophy of Science* 83.3, pp. 305–323. DOI: 10.1086/685740.

Elga, Adam and Agustín Rayo (2022). "Fragmentation and Logical Omniscience". In: *Noûs* 56.3, pp. 716–741. DOI: 10.1111/nous.12381.

Elliott, Edward (2017). "A Representation Theorem for Frequently Irrational Agents". In: *Journal of Philosophical Logic* 46.5, pp. 467–506. DOI: 10.1007/s10992-016-9408-8.

— (2022). "Comparativism and the Measurement of Partial Belief". In: *Erkenntnis* 87.6, pp. 2843–2870. DOI: 10.1007/s10670-020-00329-x.

Ellsberg, Daniel (1961). "Risk, Ambiguity, and the Savage Axioms". In: *The Quarterly Journal of Economics* 75.4, pp. 643–669. DOI: 10.2307/1884324.

Eriksson, Lina and Alan Hájek (2007). "What Are Degrees of Belief?" In: *Studia Logica* 86.2, pp. 185–213. DOI: 10.1007/s11225-007-9059-4.

Eriksson, Lina and Wlodek Rabinowicz (2013). "The Interference Problem for the Betting Interpretation of Degrees of Belief". In: *Synthese* 190.5, pp. 809–830. DOI: 10.1007/s11229-012-0187-7.

Fine, Terrence L. (1973). *Theories of Probability: An Examination of Foundations*. Academic Press.

Fishburn, Peter C. (1970). *Utility Theory for Decision Making*. Publications in Operations Research. John Wiley & Sons/Research Analysis Corporation.

— (1981). "Subjective Expected Utility: A Review of Normative Theories". In: *Theory and Decision* 13.2, pp. 139–199. DOI: 10.1007/bf00134215.

— (1986). "The Axioms of Subjective Probability". In: *Statistical Science* 1.3, pp. 335–345. DOI: 10.1214/ss/1177013611.

Freedman, David A. and Roger A. Purves (1969). "Bayes' Method for Bookies". In: *The Annals of Mathematical Statistics* 40.4, pp. 1177–1186. DOI: 10.1214/aoms/1177697494.

Gaifman, Haim and Yang Liu (2018). "A Simpler and More Realistic Subjective Decision Theory". In: *Synthese* 195.10, pp. 4205–4241. DOI: 10.1007/s11229-017-1594-6.

Gallow, J. Dmitri (2019). "Diachronic Dutch Books and Evidential Import". In: *Philosophy and Phenomenological Research* 99.1, pp. 49–80. DOI: 10.1111/phpr.12471.

Garrabrant, Scott, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor (2016). "Logical Induction". In: *arXiv preprint arXiv:1609.03543*. URL: https://arxiv.org/abs/1609.03543.

Gigerenzer, Gerd and Daniel G. Goldstein (1996). "Reasoning the Fast and Frugal Way: Models of Bounded Rationality". In: *Psychological Review* 103.4, pp. 650–669. DOI: 10.1037/0033-295x.103.4.650.

Gillies, Donald (2000). *Philosophical Theories of Probability*. Routledge.

Goldstein, Michael (1983). "The Prevision of a Prevision". In: *Journal of the American Statistical Association* 78.384, pp. 817–819. DOI: 10.1080/01621459.1983.10477026.

Golman, Russell, David Hagmann, and George Loewenstein (2017). "Information Avoidance". In: *Journal of Economic Literature* 55.1, pp. 96–135. DOI: 10.1257/jel.20151245.

Good, Irving John (1967). "On the Principle of Total Evidence". In: *British Journal for the Philosophy of Science* 17.4, pp. 319–321. DOI: 10.1093/bjps/17.4.319.

— (1974). "A Little Learning Can Be Dangerous". In: *British Journal for the Philosophy of Science* 25.4, pp. 340–342. DOI: 10.1093/bjps/25.4.340.

Greaves, Hilary and David Wallace (2005). "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility". In: *Mind* 115.459, pp. 607–632. DOI: 10.1093/mind/fzl607.

Griffiths, Thomas L., Falk Lieder, and Noah D. Goodman (2015). "Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic". In: *Topics in Cognitive Science* 7.2, pp. 217–229. DOI: https://doi.org/10.1111/tops.12142.

Gustafsson, Johan E. (2022). *Money-Pump Arguments*. Cambridge University Press.

Gyenis, Zalán and Miklós Rédei (2017). "General Properties of Bayesian Learning as Statistical Inference Determined by Conditional Expectations". In: *Review of Symbolic Logic* 10.4, pp. 719–755. DOI: 10.1017/s1755020316000502.

Hacking, Ian (1967). "Slightly More Realistic Personal Probability". In: *Philosophy of Science* 34.4, pp. 311–325. DOI: 10.1086/288169.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell (2016). "Cooperative Inverse Reinforcement Learning". In: *Advances in Neural Information Processing Systems* 29.

— (2017). "The Off-Switch Game". In: *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*. Ed. by Carles Sierra, pp. 220–227. DOI: 10.24963/ijcai.2017/32.

Hájek, Alan (1996). "'Mises Redux'—Redux: Fifteen Arguments Against Finite Frequentism". In: *Erkenntnis* 45.2-3, pp. 209–27. DOI: 10.1007/bf00276791.

— (2003). "What Conditional Probability Could Not Be". In: *Synthese* 137.3, pp. 273–323. DOI: 10.1023/b:synt.0000004904.91112.16.

Halpern, Joseph Y. (2003). *Reasoning About Uncertainty*. MIT Press.

Hammond, Peter J. (1988). "Consequentialist Foundations for Expected Utility". In: *Theory and Decision* 25.1, pp. 25–78. DOI: 10.1007/bf00129168.

Hare, Caspar (2010). "Take the Sugar". In: *Analysis* 70.2, pp. 237–247. DOI: 10.1093/analys/anp174.

Harman, Gilbert (1986). *Change in View: Principles of Reasoning*. MIT Press.

Hawthorne, James (2017). "A Logic of Comparative Support: Qualitative Conditional Probability Relations Representable by Popper Functions". In: *The Oxford Handbook of Probability and Philosophy*. Ed. by Alan Hájek and Christopher Hitchcock, pp. 277–295.

Hedden, Brian (2015). *Reasons Without Persons: Rationality, Identity, and Time*. Oxford University Press.

Hertwig, Ralph and Christoph Engel (2016). "Homo Ignorans: Deliberately Choosing Not To Know". In: *Perspectives on Psychological Science* 11.3, pp. 359–372. DOI: 10.1177/1745691616635594.

Hitchcock, Christopher (2004). "Beauty and the Bets". In: *Synthese* 139.3, pp. 405–420. DOI: 10.1023/b:synt.0000024889.29125.c0.

Holliday, Wesley H. and Thomas Icard (2013). "Measure Semantics and Qualitative Semantics for Epistemic Modals". In: *Proceedings of SALT 23*, pp. 514–534. DOI: 10.3765/salt.v23i0.2670.

Hosiasson, Janina (1931). "Why Do We Prefer Probabilities Relative to Many Data?" In: *Mind* 40.157, pp. 23–36. DOI: 10.1093/mind/xl.157.23.

Howard, Ronald (1966). "Information Value Theory". In: *IEEE Transactions on Systems Science and Cybernetics* 2.1, pp. 22–26. DOI: 10.1109/TSSC.1966.300074.

Huttegger, Simon M. (2013). "In Defense of Reflection". In: *Philosophy of Science* 80.3, pp. 413–433. DOI: 10.1086/671427.

— (2014). "Learning Experiences and the Value of Knowledge". In: *Philosophical Studies* 171.2, pp. 279–288. DOI: 10.1007/s11098-013-0267-7.

— (2017). *The Probabilistic Foundations of Rational Learning*. Cambridge University Press.

Icard, Thomas (2016). "Pragmatic Considerations on Comparative Probability". In: *Philosophy of Science* 83.3, pp. 348–370. DOI: 10.1086/685742.

— (2018). "Bayes, Bounds, and Rational Analysis". In: *Philosophy of Science* 85.1, pp. 79–101. DOI: 10.1086/694837.

— (2021). "Why Be Random?" In: *Mind* 130.517, pp. 111–139. DOI: 10.1093/mind/fzz065.

Ilin, Roman (2021). "Detection of Rare Events With Uncertain Outcomes". In: *International Journal of Approximate Reasoning* 131, pp. 252–267. DOI: 10.1016/j.ijar.2020.12.022.

Jackson, Elizabeth G. (2020). "The Relationship Between Belief and Credence". In: *Philosophy Compass* 15.6, pp. 1–13. DOI: 10.1111/phc3.12668.

Jackson, Frank and Robert Pargetter (1986). "Oughts, Options, and Actualism". In: *Philosophical Review* 95.2, pp. 233–255. DOI: 10.2307/2185591.

James, Henry ([1882] 2011). *The Portrait of a Lady*. Penguin.

Jeffrey, Richard C. (1957). "Contributions to the Theory of Inductive Probability". PhD thesis. Princeton University. URL: https://digital.library.pitt.edu/islandora/object/pitt%3A31735062223304/viewer#page/26/mode/2up.

Jeffrey, Richard C. (1990). *The Logic of Decision*. University of Chicago Press.

Jouini, Elyès and Clotilde Napp (2018). "The Impact of Health-Related Emotions on Belief Formation and Behavior". In: *Theory and Decision* 84.3, pp. 405–427. DOI: 10.1007/s11238-017-9610-3.

Joyce, James M. (1998). "A Nonpragmatic Vindication of Probabilism". In: *Philosophy of Science* 65.4, pp. 575–603. DOI: 10.1086/392661.

— (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

— (2010). "A Defense of Imprecise Credences in Inference and Decision Making". In: *Philosophical Perspectives* 24.1, pp. 281–323. DOI: 10.1111/j.1520-8583.2010.00194.x.

Kadane, Joseph B., Mark Schervish, and Teddy Seidenfeld (1996). "Reasoning to a Foregone Conclusion". In: *Journal of the American Statistical Association* 91.435, pp. 1228–1235. DOI: 10.1080/01621459.1996.10476992.

— (2008). "Is Ignorance Bliss?" In: *Journal of Philosophy* 105.1, pp. 5–36. DOI: 10.5840/jphil200810518.

Kahneman, Daniel and Amos Tversky (1973). "On the Psychology of Prediction". In: *Psychological Review* 80.4, pp. 237–251. DOI: 10.1037/h0034747.

Kant, Immanuel ([1781] 1956). *Kritik der reinen Vernunft*. First edition (A) 1781, second edition (B) 1787. Hamburg: Felix Meiner Verlag.

— ([1781] 1999). *Critique of Pure Reason*. Trans. by Paul Guyer and Allen W. Wood. First edition (A) 1781, second edition (B) 1787. Cambridge University Press.

Kelley, Mikayla and Sven Neth (2023). "Accuracy and Infinity: A Dilemma for Subjective Bayesians". In: *Synthese* 201.12, pp. 1–14. DOI: 10.1007/s11229-022-04019-9.

Kinney, David and Liam Kofi Bright (2021). "Risk Aversion and Elite-Group Ignorance". In: *Philosophy and Phenomenological Research* 106.1, pp. 35–57. DOI: 10.1111/phpr.12837.

Kolmogorov, Andrey N. (1933). *Grundlagen der Wahrscheinlichkeitsrechnung*. Springer.

Konek, Jason (2019). "Comparative Probabilities". In: *The Open Handbook of Formal Epistemology*. Ed. by Richard Pettigrew and Jonathan Weisberg. PhilPapers Foundation, pp. 267–348.

Koopman, Bernard O. (1940). "The Axioms and Algebra of Intuitive Probability". In: *Annals of Mathematics* 41.2, pp. 269–292. DOI: 10.2307/1969003.

Kraft, Charles H., John W. Pratt, and A. Seidenberg (1959). "Intuitive Probability on Finite Sets". In: *Annals of Mathematical Statistics* 30.2, pp. 408–419. DOI: 10.1214/aoms/1177706260.

Krantz, David, Duncan Luce, Patrick Suppes, and Amos Tversky (1971). *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*. New York: Academic Press.

Kreps, David (1988). *Notes on the Theory of Choice*. Underground Classics in Economics. Westview Press.

Laffont, Jean-Jacques (1989). *The Economics of Uncertainty and Information*. MIT Press.

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press.

Le Cam, Lucien (1996). "Comparison of Experiments: A Short Review". In: *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*. Vol. 30. Institute of Mathematical Statistics Lecture Notes - Monograph Series. Institute of Mathematical Statistics, pp. 127–138. DOI: 10.1214/lnms/1215453569.

Lederman, Harvey (2015). "People with Common Priors Can Agree to Disagree". In: *Review of Symbolic Logic* 8.1, pp. 11–45. DOI: 10.1017/s1755020314000380.

Leitgeb, Hannes (2014). "The Stability Theory of Belief". In: *Philosophical Review* 123.2, pp. 131–171. DOI: 10.1215/00318108-2400575.

Levi, Isaac (1987). "The Demons of Decision". In: *The Monist* 70.2, pp. 193–211. DOI: 10.5840/monist198770215.

Lewis, David (1974). "Radical Interpretation". In: *Synthese* 23, pp. 331–344. DOI: 10.1007/bf00484599.

— (1979). "Attitudes De Dicto and De Se". In: *Philosophical Review* 88.4, pp. 513–543. DOI: 10.2307/2184843.

— (1980). "A Subjectivist's Guide to Objective Chance". In: *Studies in Inductive Logic and Probability, Volume II*. Ed. by Richard C. Jeffrey. Berkeley and Los Angeles: University of California Press, pp. 263–293.

— (1981). "Causal Decision Theory". In: *Australasian Journal of Philosophy* 59.1, pp. 5–30. DOI: 10.1080/00048408112340011.

— (1999). "Why Conditionalize?" In: *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press, pp. 403–407.

Lindley, Dennis V. (1965). *Introduction to Probability and Statistics from a Bayesian Point of View. Part 2: Inference*. Cambridge University Press.

Louise, Jennie (2009). "I Won't Do It! Self-Prediction, Moral Obligation and Moral Deliberation". In: *Philosophical Studies* 146.3, pp. 327–348. DOI: 10.1007/s11098-008-9258-5.

Luce, Duncan (1967). "Sufficient Conditions for the Existence of a Finitely Additive Probability Measure". In: *The Annals of Mathematical Statistics* 38.3, pp. 780–786. DOI: 10.1214/aoms/1177698871.

MacFarlane, John (forthcoming). "Belief: What is It Good For?" In: *Erkenntnis*. DOI: 10.1007/s10670-023-00716-0.

Machina, Mark J. (1989). "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty". In: *Journal of Economic Literature* 27.4, pp. 1622–1668.

Machina, Mark J. and David Schmeidler (1992). "A More Robust Definition of Subjective Probability". In: *Econometrica* 60.4, pp. 745–780. DOI: 10.2307/2951565.

Maher, Patrick (1990). "Symptomatic Acts and the Value of Evidence in Causal Decision Theory". In: *Philosophy of Science* 57.3, pp. 479–498. DOI: 10.1086/289569.

— (1992). "Diachronic Rationality". In: *Philosophy of Science* 59.1, pp. 120–141. DOI: 10.1086/289657.

Mahtani, Anna (2012). "Diachronic Dutch Book Arguments". In: *Philosophical Review* 121.3, pp. 443–450. DOI: 10.1215/00318108-1574445.

McClennen, Edward F. (1990). *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press.

Meacham, Christopher J. G. and Jonathan Weisberg (2011). "Representation Theorems and the Foundations of Decision Theory". In: *Australasian Journal of Philosophy* 89.4, pp. 641–663. DOI: 10.1080/00048402.2010.510529.

Meehan, Alexander and Snow Zhang (forthcoming). "Kolmogorov Conditionalizers Can Be Dutch Booked (If and Only If They Are Evidentially Uncertain)". In: *Review of Symbolic Logic*, pp. 1–36. DOI: 10.1017/s1755020320000295.

Milli, Smitha, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell (2017). "Should Robots be Obedient?" In: *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*. Ed. by Carles Sierra, 4754?4760. DOI: 10.24963/ijcai.2017/662.

Misak, Cheryl (2020). *Frank Ramsey: A Sheer Excess of Powers*. Oxford University Press.

Narens, Louis and Brian Skyrms (2020). *The Pursuit of Happiness: Philosophical and Psychological Foundations of Utility*. Oxford University Press.

Neth, Sven (2019a). "Chancy Modus Ponens". In: *Analysis* 79.4, pp. 632–638. DOI: 10.1093/analys/anz022.

— (2019b). "Measuring Belief and Risk Attitude". In: *Proceedings Seventeenth Conference on Theoretical Aspects of Rationality and Knowledge*. Toulouse, France, 17-19 July 2019. Ed. by Lawrence S. Moss. Vol. 297. Electronic Proceedings in Theoretical Computer Science, pp. 252–272. DOI: 10.4204/EPTCS.297.22.

— (2023). "A Dilemma for Solomonoff Prediction". In: *Philosophy of Science* 90.2, pp. 288–306. DOI: 10.1017/psa.2022.72.

— (forthcoming). "Rational Aversion to Information". In: *British Journal for the Philosophy of Science*. DOI: 10.1086/727772.

— (forthcoming). "Better Foundations For Subjective Probability". In: *Australasian Journal of Philosophy*.

Nozick, Robert (1969). "Newcomb's Problem and Two Principles of Choice". In: *Essays in Honor of Carl G. Hempel*. Ed. by Nicholas Rescher. Reidel, pp. 114–146.

O'Donoghue, Ted and Matthew Rabin (2001). "Choice and Procrastination". In: *The Quarterly Journal of Economics* 116.1, pp. 121–60. DOI: 10.1162/003355301556365.

Oesterheld, Caspar and Vincent Conitzer (2021). "Extracting Money From Causal Decision Theorists". In: *Philosophical Quarterly* 71.4, pp. 701–716. DOI: 10.1093/pq/pqaa086.

Okasha, Samir (2016). "On the Interpretation of Decision Theory". In: *Economics & Philosophy* 32.3, pp. 409–433. DOI: 10.1017/s0266267115000346.

Oliver, Adam (2003). "A Quantitative and Qualitative Test of the Allais Paradox Using Health Outcomes". In: *Journal of Economic Psychology* 24.1, pp. 35–48. DOI: 10.1016/s0167-4870(02)00153-8.

Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury.

Osimani, Barbara (2012). "Risk Information Processing and Rational Ignoring in the Health Context". In: *The Journal of Socio-Economics* 41.2, pp. 169–179. DOI: 10.1016/j.socec.2011.10.009.

Paul, Laurie Ann (2014). *Transformative Experience*. Oxford University Press.

Pettigrew, Richard (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

— (2019). *Choosing for Changing Selves*. Oxford University Press.

— (2020). "What is Conditionalization, and Why Should We Do It?" In: *Philosophical Studies* 177.11, pp. 3427–3463. DOI: 10.1007/s11098-019-01377-y.

Pitman, Jim (1993). *Probability*. Springer.

Raiffa, Howard and Robert Schlaifer (1961). *Applied Statistical Decision Theory*. Graduate School of Business Administration, Harvard University.

Ramsey, Frank (1926). "Truth and Probability". In: *The Foundations of Mathematics and other Logical Essays*. Ed. by R.B. Braithwaite. Originally published in 1931. 1999 electronic edition. Harcourt, pp. 156–198.

— (1990). "Weight or the Value of Knowledge". In: *British Journal for the Philosophy of Science* 41.1, pp. 1–4. DOI: 10.1093/bjps/41.1.1.

Rescorla, Michael (2018). "A Dutch Book Theorem and Converse Dutch Book Theorem for Kolmogorov Conditionalization". In: *Review of Symbolic Logic* 11.4, pp. 705–735. DOI: 10.1017/s1755020317000296.

— (2023). "Reflecting on Diachronic Dutch Books". In: *Noûs* 57.3, pp. 511–538. DOI: 10.1111/nous.12409.

Resnik, Michael (1987). *Choices: An Introduction to Decision Theory*. University of Minnesota Press.

Rowell, David and Luke B. Connelly (2012). "A History of the Term 'Moral Hazard'". In: *Journal of Risk and Insurance* 79.4, pp. 1051–1075. DOI: 10.1111/j.1539-6975.2011.01448.x.

Russell, Jeffrey Sanford (forthcoming). "Fixing Stochastic Dominance". In: *British Journal for the Philosophy of Science*.

Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Russell, Stuart and Peter Norvig (2018). *Artificial Intelligence: A Modern Approach*. Third Edition. Prentice Hall.

Russell, Stuart and Devika Subramanian (1995). "Provably Bounded-Optimal Agents". In: *Journal of Artificial Intelligence Research* 2, pp. 575–609. DOI: 10.1613/jair.133.

Safra, Zvi and Eyal Sulganik (1995). "On the Nonexistence of Blackwell's Theorem-Type Results With General Preference Relations". In: *Journal of Risk and Uncertainty* 10, pp. 187–201. DOI: 10.1007/BF01207550.

Salow, Bernhard and Arif Ahmed (2019). "Don't Look Now". In: *British Journal for the Philosophy of Science* 70.2, pp. 327–350. DOI: 10.1093/bjps/axx047.

Samet, Dov (1999). "Bayesianism Without Learning". In: *Research in Economics* 53.2, pp. 227–242. DOI: 10.1006/reec.1999.0186.

Savage, Leonard J. (1967). "Difficulties in the Theory of Personal Probability". In: *Philosophy of Science* 34.4, pp. 305–310. DOI: 10.1086/288168.

— (1972). *The Foundations of Statistics*. Second Revised Edition. Wiley Publications in Statistics.

Schoenfield, Miriam (2014). "Decision Making in the Face of Parity". In: *Philosophical Perspectives* 28.1, pp. 263–277. DOI: 10.1111/phpe.12044.

Scott, Dana (1964). "Measurement Structures and Linear Inequalities". In: *Journal of Mathematical Psychology* 1.2, pp. 233–247. DOI: 10.1016/0022-2496(64)90002-1.

Sen, Amartya (2018). "The Importance of Incompleteness". In: *International Journal of Economic Theory* 14.1, pp. 9–20. DOI: 10.1111/ijet.12145.

Simon, Herbert A. (1976). "From Substantive to Procedural Rationality". In: *25 Years of Economic Theory: Retrospect and Prospect*. Ed. by T. J. Kastelein, S. K. Kuipers, W. A. Nijenhuis, and G. R. Wagenaar. Springer, pp. 65–86. DOI: 10.1007/978-1-4613-4367-7_6.

Skipper, Mattias and Jens Christian Bjerring (2022). "Bayesianism for Non-Ideal Agents". In: *Erkenntnis* 87.1, pp. 93–115. DOI: 10.1007/s10670-019-00186-3.

Skyrms, Brian (1987). "Dynamic Coherence and Probability Kinematics". In: *Philosophy of Science* 54.1, pp. 1–20. DOI: 10.1086/289350.

— (1990). "The Value of Knowledge". In: *Scientific Theories*. Ed. by C.W. Savage. Vol. 14. Minnesota Studies in the Philosophy of Science. University of Minnesota Press, pp. 245–266.

— (1993). "A Mistake in Dynamic Coherence Arguments?" In: *Philosophy of Science* 60.2, pp. 320–328. DOI: 10.1086/289735.

— (2006). "Diachronic Coherence and Radical Probabilism". In: *Philosophy of Science* 73.5, pp. 959–968. DOI: 10.1086/518815.

Smith, Holly (1976). "Dated Rightness and Moral Imperfection". In: *Philosophical Review* 85.4, pp. 449–487. DOI: 10.2307/2184275.

Soares, Nate, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky (2015). "Corrigibility". In: *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. URL: https://cdn.aaai.org/ocs/ws/ws0067/10124-45900-1-PB.pdf.

Sobel, Jordan Howard (1987). "Self-Doubts and Dutch Strategies". In: *Australasian Journal of Philosophy* 65.1, pp. 56–81. DOI: 10.1080/00048408712342771.

Spohn, Wolfgang (2012). *The Laws of Belief: Ranking Theory and its Philosophical Applications*. Oxford University Press.

Staffel, Julia (2019). *Unsettled Thoughts: A Theory of Degrees of Rationality*. Oxford University Press.

Stalnaker, Robert (1991). "The Problem of Logical Omniscience, I". In: *Synthese* 89.3, pp. 425–440. DOI: 10.1007/bf00413506.

Steele, Katie (2010). "What Are the Minimal Requirements of Rational Choice? Arguments From the Sequential-Decision Setting". In: *Theory and Decision* 68.4, pp. 463–487. DOI: 10.1007/s11238-009-9145-3.

Steele, Katie and H. Orri Stefánsson (2021). *Beyond Uncertainty: Reasoning with Unknown Possibilities*. Cambridge University Press.

Stefánsson, H. Orri (2017). "What is 'Real' in Probabilism?" In: *Australasian Journal of Philosophy* 95.3, pp. 573–587. DOI: 10.1080/00048402.2016.1224906.

Tarsney, Christian (2020). "Exceeding Expectations: Stochastic Dominance as a General Decision Theory". In: *GPI Working Paper* 3-2020. URL: https://arxiv.org/abs/1807.10895.

Teller, Paul (1973). "Conditionalization and Observation". In: *Synthese* 26.2, pp. 218–258. DOI: 10.1007/bf00873264.

Thorstad, David (2022a). "Against the Singularity Hypothesis". In: *GPI Working Paper* 19-2022. URL: https://globalprioritiesinstitute.org/wp-content/uploads/David-Thorstad-Against-the-singularity-hypothesis.pdf.

— (2022b). "Two Paradoxes of Bounded Rationality". In: *Philosophers' Imprint* 22. DOI: 10.3998/phimp.1198.

— (forthcoming). "Why Bounded Rationality (in Epistemology)?" In: *Philosophy and Phenomenological Research*. DOI: 10.1111/phpr.12978.

Titelbaum, Michael G. (2022). *Fundamentals of Bayesian Epistemology 2: Arguments, Challenges, Alternatives*. Oxford University Press.

van Fraassen, Bas C. (1984). "Belief and the Will". In: *Journal of Philosophy* 81.5, pp. 235–256. DOI: 10.2307/2026388.

— (1995). "Belief and the Problem of Ulysses and the Sirens". In: *Philosophical Studies* 77.1, pp. 7–37. DOI: 10.1007/bf00996309.

— (2023). "Reflection and Conditionalization: Comments on Michael Rescorla". In: *Noûs* 57.3, pp. 539–552. DOI: 10.1111/nous.12416.

Van Rooy, Robert (2003). "Quality and Quantity of Information Exchange". In: *Journal of Logic, Language and Information* 12, pp. 423–451. DOI: 10.1023/A:1025054901745.

Véliz, Carissa (2020). *Privacy Is Power*. Penguin (Bantam Press).

Villegas, Carlos (1964). "On Qualitative Probability σ-Algebras". In: *The Annals of Mathematical Statistics* 35.4, pp. 1787–1796. DOI: 10.1214/aoms/1177700400.

Wakker, Peter (1988). "Nonexpected Utility As Aversion of Information". In: *Journal of Behavioral Decision Making* 1.3, pp. 169–175. DOI: 10.1002/bdm.3960010305.

Weisberg, Jonathan (2007). "Conditionalization, Reflection, and Self-Knowledge". In: *Philosophical Studies* 135.2, pp. 179–197. DOI: 10.1007/s11098-007-9073-4.

Wheeler, Gregory (2021). "Less is More for Bayesians, Too." In: *Routledge Handbook on Bounded Rationality*. Ed. by Riccardo Viale. Routledge, pp. 471–483.

White, Stephen J. (2021). "Self-Prediction in Practical Reasoning: Its Role and Limits". In: *Noûs* 55.4, pp. 825–841. DOI: 10.1111/nous.12333.

Williamson, Timothy (2000). *Knowledge and its Limits*. Oxford University Press.

Yalcin, Seth (2010). "Probability Operators". In: *Philosophy Compass* 5.11, pp. 916–37. DOI: 10.1111/j.1747-9991.2010.00360.x.

Yong, Xin Hui (2023). "Risk, Rationality and (Information) Resistance: De-Rationalizing Elite-Group Ignorance". In: *Erkenntnis*. DOI: 10.1007/s10670-022-00656-1.

Zabell, Sandy L. (1989). "The Rule of Succession". In: *Erkenntnis* 31.2-3, pp. 283–321. DOI: 10.1007/BF01236567.

Zhuang, Simon and Dylan Hadfield-Menell (2020). "Consequences of Misaligned AI". In: *Advances in Neural Information Processing Systems* 33, pp. 15763–15773.

Zynda, Lyle (2000). "Representation Theorems and Realism About Degrees of Belief". In: *Philosophy of Science* 67.1, pp. 45–69. DOI: 10.1086/392761.