UNIVERSITY OF CALIFORNIA,
IRVINE

Naturalizing Decision Theory

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Daniel Alexander Herrmann

Dissertation Committee:
Chancellor's Professor Simon Huttegger, Chair
Distinguished Professor Brian Skyrms
C. H. Langford Collegiate Professor James Joyce
Chancellor's Professor Jeffrey Barrett
Assistant Professor Toby Meadows

2023

DEDICATION

To Patrick, Olga, and Jake, for giving me a strong foundation.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The members of my committee have my deepest thanks. Simon Huttegger, my committee chair, provided the optimal balance of intellectual guidance and freedom. This was essential for the development of my ideas. I am very thankful for his support these last six years. Brian Skyrms has taught me, more than anyone else, what it is to be a philosopher. His influence can be felt throughout the dissertation; when writing, I often imagined Brian as the reader. His engagement with this dissertation, as well as my development as a philosopher more broadly, has been a gift. Early in my career, the unparalleled clarity and precision of Jeffrey Barrett's work motivated me to push myself harder. I've benefited immensely from conversations with Jeff. Such conversations have the effect of making one see how things fit together, on a larger scale. They are able to do this because Jeff models genuine curiosity and intellectual flexibility. Toby Meadows probably underestimates how much of an effect he has had on my development. His depth of understanding, and ability to convey sophisticated formal material clearly, has both humbled and inspired me. Finally, I want to thank James Joyce, both as a scholar, and as a committee member. As a scholar, his work was a core example for me of how one can connect decision theory to the problems of freedom and agency that I explore in this thesis. As a committee member he has been incredibly generous with his time, attention, and encouragement. Talking philosophy with him is a real joy; Jim is a model interlocutor.

I thank my family and friends for their support throughout the program and the writing of this thesis. A few deserve special mention. Darcy Otto showed me how deep the rabbit hole goes, and how to navigate it. Gerard Rothfus has been the ideal skeptical friend; many of the ideas here were generated in discussion with him. Aydin Mohseni transformed the way I think about philosophy. Gabe Orona has pushed my ideas from many angles, and made me laugh at many moments. Simon Chen has been an unending spring of energy, interest, and good cheer. Thomas Colclough helped me to march on. Nick Cohen, Max Notarangelo, Bruce Rushing, Ronda Rushing, and Jacob VanDrunen all provided essential motivation and fellowship during writing retreats, and in life. Clara Bradley and Shasha Arani encouraged me to indulge my love of poetry. Karin Neumannová has often reminded me what it is all *for*. My parents, Patrick and Olga Herrmann, and my brother, Jake Herrmann, have my greatest thanks of all for their love and constant belief in me.

I thank the Long-Term Future Fund for its generous support of this thesis. I also thank Richard and Shura Bradley, the Notarangelos, and the good people of Green Valley Lake for providing exquisite places to write.

VITA

Daniel Alexander Herrmann

EDUCATION

Ph.D. in Philosophy **2023**
University of California, Irvine *Irvine, California*

Bachelor of Arts and Sciences **2017**
Quest University Canada *Squamish, British Columbia*

Fields of Study

Logic and Philosophy of Science

REFEREED JOURNAL PUBLICATIONS

Naturalizing Natural Salience **2023**
British Journal for the Philosophy of Science, with Jacob VanDrunen

Prediction with Expert Advice Applied to the Problem of Prediction with Expert Advice **2022**
Synthese

Sifting the Signal from the Noise **2022**
British Journal for the Philosophy of Science, with Jacob VanDrunen

Invention and Evolution of Correlated Conventions **2021**
British Journal for the Philosophy of Science, with Brian Skyrms

PAC Learning and Occam's Razor: Probably Approximately Incorrect **2020**
Philosophy of Science

ABSTRACT OF THE DISSERTATION

Naturalizing Decision Theory

By

Daniel Alexander Herrmann

Doctor of Philosophy in Philosophy

University of California, Irvine, 2023

Chancellor's Professor Simon Huttegger, Chair

This dissertation aims to *naturalize* decision theory by creating a model where an agent views herself as both a decision-maker and part of the natural world. The key contribution is a family of formal conditions that identify when an agent views herself as having control. In a slogan, an agent takes herself to control a partition if probability track her desirability. I call this approach a "desirability tracking" account of agency. I argue that this condition provides a place for individual purpose and effort, even for an agent who views herself as part of nature. I show how a desirability tracking approach allows us to chart a nuanced course between both sides of the "deliberation crowds out prediction debate" debate.

Introduction

> "Yet do thou strive; as thou art capable,
> As thou canst move about, an evident God;
> And canst oppose to each malignant hour
> Ethereal presence. . . "
> — John Keats, *Hyperion*

John Keats' *Hyperion* tells the story of the fall of the titans and the ascension of the Olympian gods. Hyperion, the titan of the sun, is the last of the titans in power. His rule is threatened by the young Apollo, the new sun god. Despite what many of Keats' contemporaries considered a remarkably promising start, Keats abandoned the poem, just as Apollo attains godhood, mid-sentence.

Many have speculated on what caused Keats to abandon *Hyperion*.[1] For the present thesis, a proposal by Bruce Miller (1965) is quite striking. According to Miller, Keats used *Hyperion* to express a philosophical puzzle, one he would

[1] See, for example, Colvin (1925), Shackford (1925), and Thorpe (1935).

need to resolve in order to write past Apollo's deification. Keats couldn't see a solution. Thus Miller's hypothesis: "Unable to solve the problem, he was unable to complete the action" (1965, p. 234).

Keats' problem stems from two conceptions of the world. The first is a world in which everything is subject to natural law. Using Oceanus[2] as his mouthpiece, Keats expresses the idea that the fate of the titans is predestined (Book II, 180-181, 211-213):

> We fall by course of Nature's law, not force
> Of thunder, or of Jove....
> So on our heels a fresh perfection treads,
> A power more strong in beauty, born of us
> And fated to excel us...

Everything that happens happens in nature, and is subject to its laws. The titans, powerful and wise as they may be, cannot escape Nature's law.

The second conception of the world is one in which there are agents who can take actions that matter. In the epigraph, for example, Coelus[3] is encouraging his son, Hyperion, to be "in the van of circumstance" (Book I, 343-344). Furthermore, Keats had plans for Apollo to be a "foreseeing God [who] will shape his actions like one".[4]

Just as Hyperion and Apollo are at odds in *Hyperion*, so too are these two

[2]The titan of the sea.

[3]The sky.

[4]This is from a letter that Keats wrote to Benjamin Robert Haydon. See page 207 of the Rollins collection of Keats' letters (2012).

worldviews. Thus, according to Miller, Keats abandoned Hyperion due to his failure to solve the following philosophical problem:

> [I]n a universe that is determined—whether toward good or evil does not affect the matter—what place is there for individual purpose and effort? (1965, p. 236)

The present thesis takes up this philosophical problem.

Instead of this problem arising in the midst of the Titanomachy, I address this problem as it arises in the context of *decision theory*.[5] Decision theory concerns the reasoning an agent carries out to make choices. In other words, it concerns deliberation. Ellery Eells gives an intuitive description:

> Deliberation is the process of *envisaging* the possible consequences of pursuing various possible courses of action and *evaluating* the merits of their possible consequences. (1982, p. 4)

The decision theory I work with in this thesis is *Bayesian*. This means it describes how a rational agent would reason, from her own perspective. Once again, Eells puts it clearly:

> Roughly, the Bayesian model says that a course of action has merit to the extent that it makes good consequences probable and that a rational person pursues a course of action that makes the best

[5] Which is no less epic.

> consequences the most probable, where the goodnesses and probabilities of the consequences are the agent's subjective assessments thereof: how true, reasonable or otherwise objectively or morally sound these assessments are is regarded as a separate question. (1982, pp. 4-5)

The goal in this context is to *naturalize* decision theory: build a model of an agent in which she can view herself as both agential and part of nature like everything else. Echoing the puzzle in *Hyperion*, we want to see if an agent who models herself as part of the universe can still find a place for her own agency.

In order to explore the subtleties that arise in this project, we will consider agents very much like Apollo and Hyperion. Though it may be strange to consider beings similar to divine intellects in a project that aims to naturalize decision theory, doing so allows us to separate two ways in which we might naturalize decision theory. The first way is to consider *bounded* agents.[6] Such agents are non-ideal reasoners, as they labour under constraints of feasibility. For example, the reasoning must be computable, or must be energetically efficient. Ultimately, if we want a full account of how agency might arise in the natural world, we will have to consider bounded agents. The second way is to characterize the reasoning of an agent who views herself as part of nature like anything else, with all of the constraints and opportunities that come with that. Instead of being concerned with feasibility, we are concerned with the tension afflicting the naturalized agent's dual self-view: as part of nature, and as agent.

[6]See, for example, Herbert Simon's *Models of Man* (1957).

In the present thesis I attempt to naturalize decision theory in the second way. In order to investigate this question, separate from the question of bounds, I consider ideally rational agents. If we interpret "foreseeing" as "probabilistically coherent", then the Keatsian version of Apollo as a "foreseeing God [who] will shape his actions like one" and a Bayesian decision theoretic agent are sesquizygotic twins.

I proceed as follows. Chapter 1 fleshes out the details of the particular kind of naturalism with which I am concerned, and introduces the core formal frameworks at play in the thesis. Chapter 2 presents the core contribution: the formal condition called *Desirability Tracking*. In a slogan, an agent takes herself to control a partition if probability track her desirability. I argue that this condition provides a place for individual purpose and effort, even for an agent who views herself as part of nature. Finally, Chapter 3 surveys the "deliberation crowds out prediction debate", and argues that Desirability Tracking allows us to chart a nuanced course between both sides of the debate.

Chapter 1

Naturalism & Decision Theory

"Wiser men than I have tried to persuade us that everything that was, is, and will be has already happened, is all of a piece, a block of manifold stuff immovably in place, and we mortals do what we do because we've already done it and thus can do no other."
— John Banville, *The Singularities*

"'Ultimately anything... may be so described. The entire universe, down to... every last particle, ray and ... event would be compressible into ... a single glyph ... single ... word.'
'Pretty long word.'
'Hopelessly so. It would take ... a universe's lifetime to articulate it. But still.'"
— Iain M. Banks, *The Hydrogen Sonata*

Chapter Summary

I introduce two conditions that any naturalistic decision theory should satisfy. The first is a richness condition: the agent's model of the world should include all of her reasoning about the world, including reasoning about herself. The second is an austerity condition: the agent's model should depend only on features of the world that the agent thinks might hold. I argue that Richard Jeffrey's decision theory satisfies the second condition, and comes close to satisfying the first. I identify a naturalistic lacuna in Jeffrey's framework: an agent's identification of her options is not presented in the framework. This sets up the project of Chapter 2.

1.1 A Familiar Objection

Decision making is an important part of our lives. When I wake up in the morning and consider how I will spend my day, I am engaged in deliberation. More gravely, if I have a serious illness and I consider which of several expensive treatments I will elect to undergo, I am engaged in deliberation.

Decision theory is the field that attempts to formalize decision making. It seeks to do so from the perspective of the decision maker, using the agent's own beliefs and values. *Descriptive* decision theory attempts to represent the decision making of actual humans. *Normative* decision theory, on the other hand, attempts to represent the decision making of ideally rational agents. The latter might be helpful for a real agent, if she is able to adjust parts of her decision making in order to better satisfy the constraints of rationality.

Another important part of our lives, not disjoint from decision making, is forming beliefs about the world. When I try to predict whether or not it will snow later in the day I am forming a belief about the world. If I am designing a microchip, then I am forming beliefs about how the chip will behave under different physical conditions. Indeed, as we have gotten better at forming detailed, accurate beliefs about the world, our ability to manipulate it through our actions has increased.

Naturalism is said in many ways in philosophy. Informally, what I mean by naturalism here is the view that agents are part of the natural world like anything else. Thus, for an agent who takes a naturalistic perspective towards *herself*, she will have beliefs about herself, and her place in the natural world. This includes beliefs about her own decision making process.

The goal of this thesis is to reconcile, within a decision theoretic framework, a naturalistic agent's view of herself as part of the world with a view of herself as a decision maker. Thus, in addition to continuing Keats' titanic work in a new context, this project lies in the Sellarsian tradition of trying to reconcile the manifest image of "man in the world" with the scientific image (Sellars (1963)). Thus, this project is a new attempt to quell the

> familiar objection that persons as responsible agents who make genuine choices between genuine alternatives, and who could on many occasions have done what in point of fact they did not do, simply can't be construed as physical systems (even broadly interpreted to include sensations and feelings) which evolve in accordance with laws of nature (statistical or non-statistical). (p.

38)

My strategy will be to go full Bayesian: instead of a metaphysically loaded sense of "genuine alternatives", the agent will have alternatives over which she is merely *epistemically* uncertain. I will argue that, when the structure of this uncertainty satisfies certain conditions, it makes sense to say that the agent views herself as an agent. Chapter 2 executes this strategy.

There is a massive and rich literature in philosophy addressing versions of this question, to which I cannot begin to do justice here.[1] This thesis focuses on this question specifically in the context of *decision theory*.

The present chapter serves three purposes. First, it clarifies a particular version of naturalism that I will use in my investigation. Second, it both introduces the decision theory framework that Richard Jeffrey provided in *The Logic of Decision* (1983), and argues that it is more naturalistic than the other standard decision theory frameworks. Finally, it argues that Jeffrey's framework still has a ways to go before it satisfies naturalism.

I proceed as follows. In §1.2 I flesh out the particular version of naturalism at play. In §1.3 I discuss how the standard approach to decision theory captures the commonsense view of ourselves as agential. Though it does a good job of capturing this view, I argue that it fails to satisfy naturalism. In § 1.4 I introduce and discuss the Bayesian approach to epistemology, which is one of the most successful and well-studied formal epistemic frameworks. I highlight

[1] As only one of many examples, Jenann Ismael considers how a situated agent might still be free, even with our best understanding of physics (2016; 2007 is also relevant). Besides focusing on how notions of freedom interact with modern physical theories, her approach differs from mine insofar as hers uses causal models, whereas I use just the agent's forward looking probabilities.

that it can model the reasoning of an agent about herself in various ways, and thus does a good job representing how an agent might conceive of herself in the scientific image. In §1.5 I introduce the decision theoretic framework that Jeffrey developed in *The Logic of Decision*, and argue that it has significant naturalistic advantages over the frameworks discussed in §1.3. In §1.6 I argue that it still fails to satisfy one of the requirements of naturalism, and briefly sketch the project of the following chapter.

1.2 First-Person Naturalism

"Naturalism" is a vague and overloaded term in philosophy. And yet it points to something useful. In this section I refine the pointing, and describe the particular kind of naturalism that I will deploy in this thesis. I also give two motivations for using the term naturalism in this way in this context. The first is that it captures a subjective version of the Sellarsian scientific image. This motivation I will weave into the description of naturalism. The second is that it makes contact with the problem of *embedded agency* that has emerged in the artificial intelligence literature. This motivation I will discuss after describing naturalism.

One more thing: the kind of naturalism here is *first-personal*, in two ways. First, it is first-personal in the sense that we are particularly concerned with her attitudes toward *herself*. Secondly, what counts as natural is agent-dependent. It only depends on how the agent *believes* the world works, not how the world *actually* works. Thus, instead of a kind of naturalism that would require the agent to model herself as consistent with how nature actu-

ally works, we only require that the agent models herself as consistent with how she *believes* nature works.

1.2.1 The Two Requirements of Naturalism

The first requirement of naturalism is that the agent models herself as part of the natural world like anything else.[2] By this, I don't require the agent to have any thick conception of natural laws. Thus, I set aside any thorny issues as to the correct status of the laws of nature. As I will make clear in §1.4, what I require is that the agent takes the same attitudes (belief and desire) towards propositions about herself as she does towards propositions about other things. This is a very weak constraint. For example, it does not require that the agent believe that the world is deterministic, or that nature is constraining in any substantial sense. What we will see is that even this meagre requirement leads to interesting tensions. There is also a richness condition: *all* of the reasoning that the agent does must be represented in the model. Putting these two together, the requirement is that (i) the agent reasons about herself the same way she reasons about anything else and (ii) all of the reasoning of the agent is represented in the model. Whereas right now this condition may be vague, the examples where it fails (in §1.3) and the examples where it succeeds (in §1.4) will make it clear. Note that this is a *richness* requirement: the model must include everything about which the agent reasons.

The second requirement is that the model must only contain things that the agent thinks are genuine possibilities. That is, if she doesn't think that some-

[2] I will give some examples of how to do this formally in §1.4 and §1.5. Here I present some motivations, and intuitions.

thing is in fact possibly part of the natural world, or if she is certain that the natural world doesn't work in a particular way, then that thing/way should not appear in the model.

This is a first-person relativized version of the "no appealing to spooky things" phrase that is sometimes used to describe naturalism. In this context, we don't forbid our agent from believing that things like gods, angels, or magic might inhabit her world. However, we do require that, if she thinks something *definitely* isn't the case, then it should not appear in our model of her thought.

Once again, while right now this may seem vague, a discussion of concrete cases in which this condition fails/succeeds in §1.5 will make this clear. Note that this is an *austerity* requirement: the model must not contain reference to anything that the agent doesn't take as a live possibility. Thus, one concrete way in which this will affect our analysis of agency is by *not* appealing to any genuine kind of counterfactuals.[3] Thus, we are looking for agency without (non-epistemic) modality.

Putting the richness and the austerity conditions together, we end up with a view of naturalism as a cognitive Goldilocks zone. The model must be rich enough to represent all the agent's reasoning about how the world might work, including reasoning about herself, but austere enough to represent no more than that.

[3]By "genuine" here I mean anything that relies on zero-probability events. For example, the imaging operation that James Joyce uses to define causal decision theory in the logic of decision (Joyce (1999)) is not a *genuine* counterfactual, since it only images on positive probability events. However, the kind of analysis of choice that Joyce (2002) uses against Levi (1997) *is* genuinely counterfactual. I disucss this in detail in chapter 3.

1.2.2 Embedded Agency

The kind of naturalism I have defined above has been getting some attention, under a different name, in parts of the AI literature. Abram Demski and Scott Garrabrant are concerned with *embedded agency* (2019). They define an embedded agent in contrast to what they call a *dualistic* agent. A dualistic agent is an agent that exists outside of its environment in some sense, and interacts with the environment only through well defined channels. These terms are perhaps somewhat foreign to philosophers. Of particular interest to us here, the dualistic agent does not need to model itself in order to take intelligent actions.[4] In order to see where these ideas are coming from, let us examine a standard way of describing an agent in computer science.

In the most popular foundational text for AI researchers, Russel and Norvig describe an agent as follows:

> An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors. A human agent has eyes, ears, and other organs for sensors, and hands, legs, mouth, and other body parts for effectors. A robotic agent substitutes cameras and infrared range finders for the sensors and various motors for the effectors. A software agent has encoded bit strings as its percepts and actions. (1995, p. 31)

[4] A formal specification of an agent that witnesses this claim is Marcus Hutter's AIXI (2004). Indeed, AIXI is an agent that only considers as live possibilities computable hypotheses, even though it itself is only semi-computable. Assuming that the environment it is interacting with is computable, Hutter shows that AIXI has many desirable optimality properties. But this assumption, especially from the naturalistic perspective here, is strong. The agent is not only not represented as part of the environment, but lives at a higher level of the arithmetic hierarchy. For more on this kind of assumption in the epistemology underlying AIXI, see Sterkenburg (2016).

This definition of an agent leads Russell and Norvig to their definition of a *rational* agent:

> For each possible percept sequence, an ideal rational agent should do whatever action is expected to maximize its performance measure, on the basis of the evidence provided by the percept sequence and whatever built-in knowledge the agent has. (1995, p. 33)

Notice that this leads quite naturally to the idea that we can describe agents as *mappings* from input percept sequences to outputs.[5] This is the model of agency that Demski and Garrabrant call *dualistic*. It is dualistic because the agent is formally separate from the environment, and only interacts with it through well-defined input output channels.

What Demski and Garrabrant are interested in in their paper, and what I am interested in here, is a different model of agency—one in which the agent and the environment are *not* formally separate, and do *not* interact via such well-defined channels. Instead of modelling the agent as somehow outside of its environment, we want to model the agent as a proper part of its environment, one with fluid boundaries (or indeed none at all!) that change over time and at different levels of description.[6] This is the kind of agent to which Demski and Garrabrant are trying to point when they use the term "embedded agent".

We see that an embedded agent is also a *naturalized* kind of agent.[7] When the

[5]They say this explicitly on the next page of the text (1995, p.34).

[6]Krakauer et al. take the first steps in developing a notion of individuality within physical systems that depends on information-theoretic properties of the system (2020). Their account has some desirable features for an embedded account of individuality: individuality is continuous (comes in degrees), exists at different levels of description, and can be nested.

[7]Indeed, in some ways, embedded agency goes beyond the type of naturalism that I

agent models its environment, and it considers itself part of its environment, then it is doing a sophisticated type of self-reasoning.

1.3 Decision Theory: The Manifest Image

We think of ourselves as agential. Though we may not have full authorship over the things that matter to us, in some situations we do get to contribute meaningfully to the evolution of the world. In these situations we have some options available to us; I may buy a lottery ticket, or not. The external world also contributes to the outcome; the number of a certain ticket is drawn to determine the winner. Together, the action and the state of the world determine an outcome.

Decision theory often represents this contribution of world and agent with an *outcome matrix*. An outcome matrix specifies an outcome for each combination of action and state of the environment. For example, suppose Hyperion is considering whether or not to fight Apollo for control of the sun. We might represent a simple version of this situation with the following matrix:

	Apollo is Stronger	Hyperion is Stronger
Fight	Imprisoned in Tartarus	Rule the Sun
Flee	Dethroned but free	Dethroned but free

discuss in this thesis, and takes into account not only the agent's reasoning about itself, but also bounds. This should also be of interest to naturalistically inclined philosophers. A model in which the agent is part of the environment fits much better our picture of how real agents in our physical world work: they are not separate from their environment, they can be influenced by the world in ways other than perception, and they can influence the world in ways other than through their actions. Perhaps, for example, the agent is heating up as it thinks about what to do, and this affects its environment in a particular way.

Here, Hyperion has two options: fight or flee. There are two possible ways the world might be, that depend on the relative strength of Apollo and Hyperion. Both Hyperion and the world jointly contribute to the outcome. The state of the world is fixed; there is nothing that Hyperion can do to change Apollo's strength. However, whether to fight or flee is up to him. This is an example of an intuitive distinction that we make when going about our lives: those things in our control, and those things out of our control.

The Bayesian perspective also requires that we specify the values and the beliefs of the decision maker. We can do the first with a *desirability matrix.*

	Apollo is Stronger	Hyperion is Stronger
Fight	-100000	100000
Flee	500	500

These numbers represent how much Hyperion desires, or values, the different outcomes. The best case is ruling the sun; the worst is to be imprisoned in Tartarus. Freedom is worth something, but is only a pale shadow of kingship.

Finally, we specify Hyperion's beliefs: how likely he takes each state of the world to be. If we have represented the problem correctly,[8] then we only need to specify two numbers: the probability of Apollo being stronger, and the probability of Hyperion being stronger.

	Apollo is Stronger	Hyperion is Stronger
Probability	0.97	0.03

[8]This requires that the states be chosen such that states and acts are appropriately independent of each other. What kind of independence—evidential or causal—is the subject of much spilled ink.

These attitudes fit our commonsense view of decision making. When making decision, we can consider what we might do, consider how the world might be, and try to act so as, roughly, to make good outcomes likely, and bad outcomes unlikely. The Bayesian picture formalizes this; a rational agent chooses the act that maximizes the expected desirability. To calculate this for a given act, we take the probability of each state, multiply it by the desirability of the outcome that the act and state jointly produce, and then add. For example, the expected utility of Hyperion's different acts are as follows:

$$\text{EU}(\text{Fight}) = 0.97 * -100000 + 0.03 * 100000 = -94000$$

$$\text{EU}(\text{Flee}) = 0.97 * 500 + 0.03 * 500 = 500$$

Since $500 > -94000$, Hyperion should flee.

Leonard Savage developed this basic picture into a sophisticated and fruitful theory of decision (1972). Savage's theory is still the core decision theory in economics.

Many aspects of this basic picture are left unchanged. There are three basic types of objects: states, acts, and outcomes. Like above, states are features of the world that help to determine outcomes. These are the objects of belief for the agent. Outcomes are results from the interaction of the agent and the world: these are objects of desire for the agent. Acts are *functions* from states to outcomes. That is, an act determines a unique outcome for each state. Acts are the objects of choice for the agent.

Instead of assuming that the agent has basic desirability and probability judg-

ments, Savage showed how to extract such judgments from an agent's *preference* ordering over all possible acts. We can think Savage as providing a way to measure belief and desire from (very idealized) behavior.

Here is an intuitive picture. Think again of a decision matrix, but this time a titanic one. Each column in the matrix corresponds to one possible state of the world. Furthermore, there is a column for each state of the world the agent considers. This will be a very wide matrix.

It will also be a very tall matrix. Each row in the matrix corresponds to one possible act. Recall that acts are functions from states to outcomes. In order to carry out this measurement procedure, Savage requires that *every* function from states to outcomes is somewhere in this matrix. This is called the *Rectangular Field Assumption* (Broome (2017)). This generates a massive number of rows. For example, if there are only 10 possible states and 10 possible consequences, then there will already be $10^{10} = 10000000000$ rows![9]

Now that we have this matrix, we can understand Savage's measurement procedure. Instead of basic probabilities and desirabilities, Savage considers an agent who has a preference ranking over all the acts in the matrix. For example, if f and g are two acts, the agent might prefer f to g, written $f \succ g$.

Instead of thinking of this as a concrete decision problem in which only one act will be chosen, like Hyperion's problem above, we instead think of this as a recipe for generating the choice of the agent in any given decision problem. Any such decision problem will be given by a subset of the set of all possible acts. The behavioural data then corresponds to observing how the agent chooses in

[9]In fact, the set of possible functions in Savage will be the size of the continuum, or larger.

every possible decision problem.

What Savage shows is quite remarkable. If the agent's preference ranking over all possible acts satisfies certain conditions,[10] then we can represent the agent's preference ranking with a unique probability measure on states and a semi-unique desirability function on outcomes, where

$$f \succ g \iff \mathrm{EU}(f) > \mathrm{EU}(g)$$

This provides a firmer foundation for the Bayesian claim that the rational agent is one who maximizes expected utility. If an agent's preferences satisfy certain plausible rationality requirements, then we can think of her as maximising expected desirability.

Savage's theory has many virtues. It extends our standard, manifest view of decision making, in which our acts and the world jointly determine outcomes. It provides a firm foundation for expected desirability. It is flexible enough to represent many different decision problems.

How does it fare on our two naturalism requirements? Unfortunately, it fails them both.

The first condition was a richness condition. The model of decision making should include all of the reasoning the agent does, and the agent should have the same attitudes towards propositions about herself as she does towards propositions about other parts of the world. This second aspect is clearly

[10]Some of these conditions are *rationality* requirements, that are supposed to capture rational choice in an intuitive way. Others are *structural* requirements, that are needed for the mathematics.

violated. Savage's agent assigns probabilities to states;[11] yet Savage's agent does not assign probabilities to her own acts.[12]

The first part of the richness condition is also violated. Recall the idea of the Savage preference ranking as a *recipe* for generating decisions in different decision problems. The idea is that, when faced with a subset of all acts as the ones available for choice, the agent should/does pick the one that ranks highest in the preference ordering. However, a part of that story that is not represented in the model is how the agent comes to believe she has that subset of acts available. This can matter, as it might be the case that coming to believe that some act is available to you might radically reshape your theory of the world.

An example will help make this clear. This example also illustrates how the austerity condition is violated. Consider an agent who has as possible states[13] whether or not it will rain the next day. Suppose further that possible consequences that matter to her are whether or not there will be a nuclear war. Now, by the Rectangular Field Assumption, an agent will have in her preference ranking an act that maps every state where it rains to a nuclear war, and every state where it does not rain to peace.

Consider, now, that this agent has to have a preference between this and any other act. Jeffrey writes the following:

[11] Really, Savage's agent assigns probablities to sets of states.

[12] Furthermore, Savage's agent does not assign unconditional probabilities to consequences. She only assigns probabilities to consequences, *conditional* on acts. Similarly, Savage's agent does not assign desirability to states of the world, even though intuitively she may desire certain states of the world to be the case. She does, however, assign desirability to states of the world conditional on a particular act. In general, this kind of infelicity arising from the tripartite structure of Savage's framework clashes with naturalism.

[13] Really, events.

> However, for the agent to consider that this [act] might be in effect would require him so radically to revise his view of the causes of war and weather as to make nonsense of whatever judgment he might offer.[14] (1983, p. 157)

Jeffrey is expressing a naturalistic concern here. We can view the Rectangular Field Assumption as leading to either a failure of richness or a failure of austerity. If it is the case that the model does not include the reasoning the agent does when considering this act as available, which as Jeffrey points out might be fairly dramatic, then it fails the richness requirement. However, if the model does include this reasoning, then it includes too much, for the reasons Jeffrey points out: the judgement lies outside of the agent's picture of the natural world, for she doesn't actually believe every possible function from states to outcomes to be possible, and thus becomes meaningless.

Thus we see that Savage's decision theory is non-naturalistic.[15] It includes too much due to the Rectangular Field Assumption. In includes too little because the agent doesn't have beliefs about her own acts, and because it doesn't represent how the agent's view changes as she learns that different acts are available to her.

As we will see in §1.5, Jeffrey's decision theory has naturalistic advantages over Savage's. In particular, the set of objects over which agents have preferences

[14] Jeffrey here is actually discussing gambles in the context of the theory Frank Ramsey develops in *Truth and Probability* (1931). However, the issue is the same in Ramsey and Savage. The richness condition in the set of acts/gambles leads to fantastic objects in the preference ranking, that violate the agent's theory of the world. We will see in §1.5 how Jeffrey's theory gets around this.

[15] Everything said here also holds for Ramsey's decision theory.

in Jeffrey's theory is, in a sense, much smaller.[16]

Before we move to Jeffrey's theory we will first build up some basic tools of Bayesian epistemology. This will help us understand how an agent can reason about propositions that describe herself. It will also help us understand the basic structure of Jeffrey's theory, which differs from that of Savage's.

1.4 Bayesian Epistemology: The Scientific Image

Recall that one of the core aspects of naturalism is that the agent take the same attitudes towards propositions about herself as she does towards propositions about other things. As far as other types of naturalism go, this is a weak requirement. We do not require that the agent's view on things and herself be compatible with our best theories of physics, for example. Ultimately, we *will* want such an account. But naturalism here is first-personal. All that we require is that the agent's view of herself is compatible with *her* theory of the world.

The Bayesian approach to epistemology will form the basis of our model of the agent's attitudes towards the world. I use the Bayesian approach here for three reasons. The first is that it is an incredibly successful and well-studied formalization of epistemology, and is supported from a number of different

[16]For the specialist: instead of the objects of preferences being all possible functions, which represent all possible causal connections between states and outcomes, the objects of preference are an algebra of propositions, and thus represent only the logical connections between propositions. I will go over this in detail in §1.5.

angles.[17] The second is that it is continuous with Bayesian decision theory, and thus allows us to explore the question of how an agent's view of herself as decision maker can cohere with a view of herself as part of the world.[18] Finally, it is a radically subjective epistemology, that describes the agent's reasoning from her point of view. This is desirable for us given that our naturalism is first-personal.

Bayesian epistemology relies on a type of mathematical object called a *probability space*. A probability space is an ordered set $\langle \Omega, \mathcal{A}, P \rangle$. In Bayesian epistemology this object represents the epistemic attitudes of the agent. Ω is a set of personally possible worlds. These represent all of the possible ways the world could be, *from the agent's point of view*.[19] For example, if we are representing an agent who has beliefs exclusively about the outcomes of a die roll, then her set of possible worlds might be

[17] Jonathan Weisberg provides a nice summary of justifications for the Bayesian point of view (2011).

[18] Indeed, one of the main approaches to justifying Bayesian epistemology uses the kind of representation theorem discussed in §1.3.

[19] I highlight here that the set of possible worlds is personally possible for two reasons. The first is that I want my account to be thoroughgoingly first-personal. The second is that there has been great concern among philosophers about the problem of *logical omniscience*. Supposedly, being a Bayesian commits one to knowing all logical and mathematical facts. However, this is because the possible worlds that some philosophers imagine underlie the Bayesian framework are metaphysically possible worlds, instead of personally possible. This begs the question. Instead of choosing the underlying set of worlds such that it forces logical omniscience, one should choose it such that it represents what the agent herself considers possible, at her current state of knowledge. An example of how to do this for decision-relevant logical uncertainty is given in Lipman (1991). The idea is to construct the agent's sample space such that any fact of logic/math about which the agent is uncertain is represented by some proposition in the agent's algebra. If one likes, one can think of this kind of approach as one where we allow an agent to have "logically impossible" worlds in her algebra. Hacking argues for basically the same idea (1967). Pettigrew makes a similar point (2021). This approach is similar to Hintikka's use of impossible worlds to deal with logical omniscience in his account of knowledge (1979). This is important for our purposes here, because the account of deliberation I develop will have computations as one possible type of learning event.

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

The next component, $\mathcal{A}$, is an *algebra*.[20] This algebra consists of subsets of Ω. We call each element of the algebra an *event* or a *proposition*. These are the objects over which agents have beliefs. For example, what we would describe in English as "the die came up even" and "the die came up greater than 4" would be represented with the sets

$$\{2, 4, 6\} \text{ and } \{5, 6\}$$

respectively. These are both events.

Events can be much more complicated than mere outcomes of die rolls. They can express states of affairs such as "a barn owl will catch a mouse in Irvine today" and "I drink a glass of Chianti".

In addition, we can combine propositions using the connectives of sentential logic to yield new propositions. For example, if A and B are propositions, then so are $\neg A$, $A \vee B$, $A \wedge B$, and $A \rightarrow B := \neg A \vee B$. Working with propositions as sets, negation corresponds to complement, disjunction with union, and conjunction with intersection. Given this correspondence, we will move comfortably between the operations of logic and the operations of set theory.[21]

[20] For the formally inclined, this is usually required to be a σ-algebra.

[21] For example, "the die came up even and the die came up greater than 4" would be $\{2, 4, 6\} \cap \{5, 6\} = \{6\}$.

Not only can we combine propositions in this way, but we require that $\mathcal{A}$ be *closed* under these operations. This is because we think of $\mathcal{A}$ as capturing the conceptual resources of the agent. The idea is that if she can consider certain propositions, then she can consider any logical combination of those propositions. This motivates the closure conditions: if the agent can consider A, then she should be able to consider $\neg A$. Thus, if $A \in \mathcal{A}$ then $\neg A \in \mathcal{A}$. A similar motivation holds for conjunction and disjunction. Finally, we require that $\mathcal{A}$ contain a necessary proposition, the unit T, and the impossible proposition, the zero F. Thinking in terms of sets, T is the set of all possible worlds, and F is the empty set. Thus, we have that $F = \neg T$. For every proposition $A \in \mathfrak{A}$, $T = A \vee \neg A$. Since $\mathcal{A}$ satisfies these conditions, this means that $\mathcal{A}$ is a Boolean algebra.

Finally, P is a probability function defined on $\mathcal{A}$.[22] $P(A)$ represents the degree of belief that the agent has in each proposition $A \in \mathcal{A}$. Because P is a probability function is satisfies the following constraints:

The value of P is bounded between 0 and 1 (inclusive). P must assign a value of 1 to Ω; the agent is certain that *something* happens. If two events A and B are incompatible ($A \cap B = \emptyset$), then $P(A \cup B) = P(A) + P(B)$. Requiring that P is a probability reflects the fact that an agent should be internally coherent.

This is the core of the Bayesian framework. Often we use a probability space to represent an agent's reasoning in a particular context. For example, if we want to represent the reasoning of an agent predicting the outcomes of coin tosses, we might use a probability distribution over Cantor space to do so.[23]

[22] P is usually required to be countably additive.

[23] Cantor space here is the set of all infinite binary sequences, with the intended interpretation that "1" means "heads" and "0" means tails. It also implicitly includes a σ-algebra

In our context, however, we are interested in the agent's entire theory of the world. Thus, in contrast to most applications of Bayesian epistemology, we imagine that Ω contains every world conceivable to the agent and $\mathcal{A}$ every set of these towards which the agent has attitudes. Thus P represents the entirety of the agent's epistemic attitudes towards every proposition she can consider. In general, we should expect $\mathcal{A}$ to be *massive*. Such an idealized epistemology is for the brilliant Hyperion and the prophetic Apollo: not for us mere mortals.

Considering such an idealized model allows us to bring into stark relief our central concern: the agent's self-conception. If the agent considers herself as part of the world, and the algebra contains every proposition the agent can consider, then the algebra will have propositions that refer to the agent.

This is actually fairly typical in formal epistemology. Consider, for example, the principle of *reflection*. Informally, reflection says that an agent should defer to her future degrees of belief. Simon Huttegger gives a number of different ways to make this precise (2013). Here is one:

> An agent's current degree of belief in event A given that her anticipated future degree of belief $\mathbb{P}_f[A] = r$ should be equal to r with probability 1, whenever the event $\mathbb{P}_f[A] = r$ has positive probability. (2013, p. 414)

Notice that this condition requires that there be an event, "$\mathbb{P}_f[A] = r$", in the agent's algebra. This is a proposition about the agent's future degrees of belief. Thinking in terms of an algebra of sets, this would be the set of

that is a subset of the powerset of all infinite binary sequences.

possible worlds at which the agent has degree of belief r in proposition A at a certain time. Indeed, when Huttegger builds the probability space, he ensures that the space is rich enough to contain instances of this kind of auto-epistemic event, so that he can lay a careful foundation for reflection. Thus, not only are propositions about the agent possible and typical in Bayesian epistemology; they are also theoretically fruitful.

Briefly summarizing this section, we see that Bayesian epistemology, and its key mathematical device of a probability space, provide a very general way of representing the agent's view of the world. This, then, is the agent's personal *scientific image* of things. Propositions about an agent's own mental states play a key role in stating and grounding Bayesian principles, such as reflection. When an agent has such auto-epistemic propositions in her algebra, they represent the scientific image she has of herself.

Thus we have our two ingredients. Decision theory captures the manifest image of decision making. The auto-epistemic propositions capture the scientific image an agent has of herself. We want to see how far we can push these together. Now we turn to an example of a theory that pushes them fairly close.

1.5 The Logic of Decision

In this section I introduce the theory Jeffrey developed in *The Logic of Decision* (1983). Since this will be the core framework I work with in the thesis, I will spend some time working through the details. I show that it has naturalistic

advantages over the Savage framework.

1.5.1 The Theory

In Savage's theory, an agent has preferences over a set of acts. In Bayesian epistemology, the agent has probabilities defined on an algebra of propositions. Jeffrey's theory puts the ideas together. In the Jeffrey-Bolker theory, an agent has preferences, desirabilities, and probabilities, all defined over the same set of objects: an algebra of propositions. Call this algebra $\mathfrak{A}$. This already differs from Savage's theory, in which probabilities are defined over (an algebra of) states, and utilities are defined on consequences (and acts), which are formally separate objects.

Just as in §1.4, $\mathfrak{A}$ must be closed under the (countable) operations of logic. In addition to this requirement, there are two more conditions that the algebra must satisfy. First, we require that $\mathfrak{A}$ is atomless. This means that, for all $A \in \mathfrak{A}$ such that $A \neq F$, there is some $B \in \mathfrak{A}$ such that $B \rightarrow A$ and $F \neq B \neq A$. What this means is that for any proposition A, we can always break it down into two further propositions B and $\neg B$. This is why the algebra is *atomless*; an atom would be a proposition that could not be further divided. Atomlessness is required so that we can think of all desirability as expected desirability.[24]

Finally, the algebra must be *complete*, which means that every subset of the

[24]Sometimes the philosophical implications of this feature of Jeffrey's framework are oversold (for example, in section 3.2 of Steele and Stefánsson (2020)). It is entirely possible recover a utility function defined not on the members of the algebra, but on the possible worlds. This is the little u Jeffrey writes in equation (2) in *A Note on the Kinematics of Preference* (1977).

algebra has both a *supremum* and an *infimum*. The supremum of a set of propositions $\mathcal{C} \subset \mathfrak{A}$ is a proposition S such that for all $A \in \mathcal{C}, A \to S$, and for all S' such that S' also has this property, $S \to S'$. Similarly for the infimum, but with the direction of implication reversed. Suprema and infima are unique. Completeness is required so that we can state a continuity constraint on the agent's preferences.

Now that we have the objects of preference defined, we can talk about the preferences themselves. Similar to Savage, the goal is to start off with an agent's qualitative preference ordering over propositions, and show that, when this ordering satisfies certain conditions, then the preference ordering can be represented as arising from a probability and a utility function that jointly determine preferences via expected utility maximization.

The preference relation $\succeq$ is defined on $\mathfrak{A}' := \mathfrak{A} - F$. We interpret $A \succeq B$ as saying "B is not preferred to A" or equivalently, "A is at least as preferred as B". We require the preference ordering to be a complete preorder on $\mathfrak{A}'$.[25] We also require that the preferences satisfy a certain technical continuity condition.[26] We write $A \succ B$ iff $A \succeq B$ and $B \not\succeq A$. This means that the agent strictly prefers A to B. We write $A \approx B$ iff $A \succeq B$ and $B \succeq A$; this means that the agent is indifferent between A and B.

We can now state the core two axioms of the Jeffrey-Bolker framework: averaging and impartiality. These are the following conditions:[27]

[25] This means that $\succeq$ is reflexive and transitive.

[26] Specifically, let $\mathcal{A} = \{A_1, A_2, \ldots\} \subseteq \mathfrak{A}$ be a sequence of propositions such that $A_n \subseteq A_{n+1}$, for all n. Then, if A^* is the supremum (infimum) of $\mathcal{A}$ and $B \succ A^* \succ C$, then there is some N such that $B \succ A_n \succ C, \forall n \geq (\leq) N$.

[27] In these axioms, all propositions are assumed to be in $\mathfrak{A}'$.

1. (Averaging) If $A \wedge B = F$, then

   (a) if $A \succ B$, then $A \succ A \vee B \succ B$, and

   (b) if $A \approx B$, then $A \approx A \vee B \approx B$.

2. (Impartiality) Whenever $A \wedge B = F$ and $A \approx B$, then, if $A \vee C \approx B \vee C$ for some C such that $A \wedge C = B \wedge C = F$, and $C \not\approx A$, then $A \vee C \approx B \vee C$ for every such C.

Averaging ensures that the disjunction of two propositions lies between the two propositions. For example, if you prefer visiting the Museum of Jurassic Technology to visiting the Getty Museum, then a particular gamble between both visits should be dispreferred to surely visiting the Museum of Jurassic Technology, and preferred to surely visiting the Getty Museum.[28]

Impartiality basically gives a way to test that the agent views two disjoint propositions (A and B) as equally probable. Let us walk through how this works. Suppose A and B are disjoint (meaning $A \wedge B = F$), and suppose that the agent is indifferent between them ($A \approx B$). In the background, we are imagining that the preference ranking has arisen by the principle of maximizing expected utility.[29] Then, the agent will be indifferent between $A \vee C$ and $B \vee C$ only in the case where A and B are equiprobable. For, suppose not. For concreteness, imagine that $C \succ A$, and that B is more probable than A. Then the agent will prefer $A \vee C$ to $B \vee C$, since this gives

[28] If you squint hard enough, then Averaging is somewhat similar to the irrelevance of independent alternatives axioms present in other utility theories (for example, in Von Neumann and Morgenstern (1953)), and Savage's Sure-Thing Principle. But they are also very different. I will discuss this difference in more detail in §1.5.2.

[29] This *is* kind of cheating, since we are very clearly building into the axioms what we want to get out. Jeffrey makes this point on page 147 of *The Logic of Decision* (1983).

the agent a higher probability of getting the more desired outcome, C. This is how this condition works as a kind of test for equiprobability. The axiom says that this test will come out the same way, *no matter which* test proposition C is used.

Now that we have all of the conditions we are almost ready to state the main theorem of the theory. We just need one additional piece of terminology: that of a *signed measure*. A signed measure is a generalization of the notion of *measure* which is allowed to take on negative values.[30] Using a signed measure makes sense in the context of the theorem, since it will play a role in defining *desirability*, which is allowed to be negative.

Again, the goal is to show that we can think of an agent whose qualitative preferences satisfy these axioms *as if* she were an expected utility maximizer of some kind. We call a preference ordering that satisfies these conditions *coherent*. The following theorem shows us this:

Bolker's Existence Theorem. Let $\mathfrak{A}$ be a complete atomless Boolean algebra, and let $\succeq$ be a coherent preference ordering on $\mathfrak{A}'$. Then there exists a probability measure P defined on $\mathfrak{A}$ and a signed measure v on $\mathfrak{A}$ such that, for all A and B in $\mathfrak{A}'$:

$$A \succeq B \text{ iff } U(A) \geq U(B)$$

where

$$U(A) = \frac{v(A)}{P(A)}, \forall A \in \mathfrak{A}'.$$

[30] A *measure* is a function on an algebra that assigns 0 to the bottom element, is non-negative, and is countably additive.

It follows that, for any $A \in \mathfrak{A}'$ and countable partition $\mathcal{S} = \{S_i\}_{i \in I}$ of A, we have that we can calculate the utility of A as follows:

$$U(A) = \sum_{i \in I} U(A \wedge S_i) P(S_i | A).$$

This makes clear that the representation is in fact one of expected utility.

Finally, a brief note on the *uniqueness* of the representation. In general, the representation is not uniqune. The following theorem makes this precise:

Bolker's Uniqueness Theorem. Let P, P' be probability measures on $\mathfrak{A}$ and let v, v' be signed measures on $\mathfrak{A}$. Then the pair P', v' represents the same preference order as P, v iff:

$$v' = av + bP, \text{ and } P' = cv + dP$$

where $ad - bc > 0$, $cv(T) + d = 1$, and for all $A \in \mathfrak{A}'$ $cv(A) + dP(A) > 0$.

This transformation of P and v to P' and v' induces the following shift in U:

$$U' = \frac{v'}{P'} = \frac{av + bP}{cv + dP} = \frac{aU + b}{cU + d}.$$

Thus, in the Jeffrey-Bolker theory, the probability function is also non-unique. The account of decision making I give in this thesis will require that we work

with a unique probability.[31] There are a number of conditions we can impose on the representation to yield a unique probability.

The first is to require that U be unbounded above and below. In that case, the probability function is unique, and the utility function is unique up to positive linear transformation. Thus, utility is an interval scale.

A second strategy is to introduce constraints on *comparative belief*. When these constraints are satisfied, then, once again, the probability function is unique, and the utility function is unique up to positive linear transformation. Joyce (1999) extends the work of Villegas (1964) and makes this precise. Similarly, Ahmed (2014) shows that, if we assume that an agent has primitive "equal confidence" judgments about propositions, and we assume some conditions on how this works, then we also get a unique probability and an interval scale for utility.

In the rest of this thesis I assume that we have conditions that yield a unique probability and an interval scale for utility.

1.5.2 Naturalism in the Logic of Decision

Now that we have Jeffrey's theory on the table, we can see that it has naturalistic advantages over Savage. The most obvious advantage it that it allows agents to have attitudes toward propositions about herself, just like anything else. This is the same as the Bayesian framework discussed in § 1.4. Propositions about the agent's degrees of belief, desires, and acts can all be repre-

[31] The extent to which my account actually requires a unique probability is an interesting question, and one I plan to address in future work. For now I assume it for convenience.

sented in Jeffrey's framework. This means that, if the agent's algebra has such propositions, and the algebra of the agents we are working with does, then the agent will have probabilities and desires defined over propositions about herself. Furthermore, since such propositions are also all part of the same algebra as propositions describing the other parts of the natural world, this means that, since $\mathfrak{A}$ is closed under logic, the agent can consider propositions that are about herself and other parts of the world. This is of course dramatically different from a framework like Savage's, in which the agent's acts are formally separate from propositions about the rest of the world.[32]

There is another, subtle, advantage that Jeffrey's framework has over Savage's. While the first advantage was a richness one—a Jeffrey agent can have probabilities and desires over propositions about herself—the second advantage is an austerity advantage. The advantage lies in the domain of the preference relation. Recall that Savage has the Rectangular Field Assumption: the agent's preference ranking must include every possible function from states to consequences. If we think of these acts as describing *causal* connections between states and acts, then we can understand the Rectangular Field Assumption as a certain causal condition. In Jeffrey's theory, on the other hand, the agent's preference ranking only includes every proposition in $\mathfrak{A}'$. Jeffrey writes the following about the difference between his theory and Ramsey's decision theory, but the exact same difference holds between Jeffrey and Savage's decision

[32]Some have taken this property of the Savage framework to be a feature, not a bug. For example, Spohn argues that any adequate decision theory must *not* have an agent assign probabilities to her own acts (1977). This position has been argued for further—see, for example, Levi (1993) and Levi (2007). On the other side of the debate are, for example, Joyce (2002), Rabinowicz (2002), and Hájek (2016). However, at least on the surface, from a naturalistic position, denying the possibility of act probabilities induces such a dualistic perspective of oneself that is deeply unsatisfactory. In the third chapter, after developing my account of action, I consider the act-probabilities debate. It turns out that the account I develop walks the line between both camps.

theory:

> In these terms, the primary difference between Ramsey's theory and ours is the difference between the *causal* operation and the *logical* operations. Ramsey measures his agent's desirability and probability assignments by presenting him with a bewildering variety of possible causal connections between propositions and consequences, some of which are wildly at variance with the agent's notions of how things happen in the world. We perform the corresponding measurements by presenting our agent with a less bewildering variety of entities: with all possible combinations that can be formed by applying the logical operations *not*, *and*, and *and/or* to any propositions between which he has preferences or between which he is indifferent.... [But in Ramsey's theory] to ask the agent to locate [an act that maps Y to X and $\neg Y$ to Z] in his preference ranking when X, Y, and Z are the propositions that (X) there will be a thermonuclear war next week, (Y) this coin will land head up when I toss it, and (Z) there will be fine weather next week, is not to invite him to take pains in the interest of clarity and self-knowledge. To the extent that he can bring himself to consider the gamble seriously, he must entertain alarming and bizarre hypotheses about the person who is offering the gamble: hypotheses that he can only entertain by altering his sober judgements about the causes of war and weather, and thereby altering the very probability assignments which the method reports to measure. (1983, pp. 159-160)

This is a particularly striking passage. Jeffrey is deeply sensitive to the concern that Ramsey's theory (and Savage's) requires the agent to entertain objects that are "wildly" incompatible with how the world actually works. This is a failure to satisfy the austerity condition of naturalism. Thus, when Jeffrey writes, "I take it to be the principal virtue of the present theory, that it makes no use of the notion of a gamble or of any other causal notion" (1983, p. 157), his reason is deeply naturalistic.[33]

The austerity of Jeffrey's framework makes a real difference in how the framework functions. Taking a closer look at Jeffrey's example, consider two acts in Savage: f and g, defined as $f(y) = X, f(y') = Z, g(y) = X, g(y') = V, \forall y \in Y, y' \notin Y$, and X, Y, and Z as in the Jeffrey passage above, and V denoting the consequence that the agent will receive a package in the mail that contains one billion CAD. Suppose, furthermore, that the agent prefers receiving the money to fine weather next week. Then, in Savage's framework, it must be the case that the agent prefers g to f.[34]

Though this is a theorem of Savage's framework,[35] in the von Neumann-

[33]Ethan Bolker, the mathematician responsible for much of the mathematics underlying the system in *The Logic of Decision*, was also motivated by austerity concerns. For example, he writes,

> The 'Bolker objection' (which could just as well have been named the Jeffrey objection) says that it is unreasonable to ask a decision maker to express preferences about events or lotteries he feels cannot occur. (p. 80, 1974)

Indeed, this is what he leads him to have a preference for a strictly positive probability measure:

> We need not worry about zero denominators since the choice of [states] and [the set of events] is our subject's; he simply will not consider to him any states which seem impossible. (p. 337, 1967)

[34]Really it also has to be the case that $\neg Y$ is not null. Then $g \succ f$ follows from Theorem 2 of Savage, taking B to be the set of all states (1972, p. 24).

[35]Indeed, it only requires the first two postulates.

Morgenstern (VNM) framework for expected utility this is one of the axioms. In the VNM framework the agent has preferences over a set of gambles, where gambles are arbitrary probability distributions over a finite set of outcomes. If the agent's preferences over all such gambles satisfy certain conditions, then we can recover a quasi-unique utility function over consequences, such that the agent's preferences go by expected utility (Von Neumann and Morgenstern (1953)). One of the conditions is *Independence*:

Independence. For any N, L, and M in the set of gambles, and any $p \in (0, 1]$, $L \preceq M$ iff $pL + (1 - p)N \preceq pM + (1 - p)N$.

Given that this is one of the axioms of the VNM theory, it is clearly intuitive for most people. The idea is that, if you know that you will only get one of the consequences, and one of the consequences is the same across gambles (N), then the preference between the gambles must agree with the preference between the parts of the gambles that differ (L vs. M). It may be surprising, then, that a seemingly analogous condition can fail in Jeffrey's theory. In Jeffrey's theory, we might expect something like the following to hold:

J-Independence. For any pairwise incompatible propositions N, L, and M in $\mathfrak{A}$, $L \preceq M$ iff $L \vee N \preceq M \vee N$.

In this condition "$\vee$" is playing the roll of forming gambles over different propositions. However, it can fail. Consider again Jeffrey's example. Whereas in Savage we had f and g, with Jeffrey, if we assume that the three consequences are all pairwise disjoint, we would have the following propositions as surrogates for f and g:

$$X \vee Z, X \vee V.$$

Now, we supposed in Savage's case that the agent preferred receiving the billion CAD to fine weather, so let us do the same here. Thus, we have $V \succ Z$. Again, we might suspect that this implies $X \vee V \succ X \vee Z$. However, in this case this might plausibly fail. For, suppose that the agent considers V to be dramatically less probable than Z. This is, in fact, my current belief in my own situation. In Jeffrey's framework, when taking disjoint unions, *propositions carry their own probability.* Thus, when we take disjunctions, if the probability of X is sufficiently high, and the desirability of X is sufficiently low, the preference can flip. In this example, I certainly prefer receiving a billion CAD to there being fine weather next week. However, I prefer the proposition "there will be a thermonuclear war next week or there will be fine weather next week" to the proposition "there will be a thermonuclear war next week or I will receive a billion CAD". This is because, conditional on the former proposition most of my probability mass lies on fine weather, whereas conditional on the former proposition most of my probability mass lies on thermonuclear war.

What allows Independence to hold in Savage and VNM style theories is that the probabilities with which different consequences obtain have nothing to do with the consequences themselves, and thus can be made identical across gambles. This is what p does in Independence, and what the definition of f and g do in the Savage example. In VNM the probabilities over consequences are stipulated. In Savage, the probabilities over consequences are given by

probability over states, as induced by arbitrary functions. In both cases, the agent is considering possibilities that "are wildly at variance with the agent's notions of how things happen in the world" (Jeffrey (1983), p. 159).

Working through the failure of J-Independence shows us how naturalism manifests in Jeffrey's theory, and also that it can make a substantial difference. I agree with Jeffrey that the austerity of his framework is its main advantage. This, in addition to its ability to support an agent reasoning about her own acts, make it a quite naturalized framework. Despite this success, there remains a naturalist lacuna. It is to this that I now turn.

1.6 Exogenous Options

Our species of naturalism requires that the model represent all of the reasoning of the agent, including reasoning about the agent itself. It turns out that, while Jeffrey does better than Savage at including such reasoning, Jeffrey is still missing something. Furthermore, this is something that Jeffrey himself thought was an important part of the story:

> Deliberation–deciding what to do–is a matter not only of clarifying your preferences, but of identifying your options. (1977, p. 137)

Jeffrey wants a theory that not only represents the agent's preferences, but also one that lets the agent identify her options. And yet, there is nothing in his theory that plays this role.

To see this, let us briefly turn back to Savage. Recall in §1.3 that we can think

of the preference ordering that a Savage agent has over all the acts as providing a recipe for making decisions. Since the agent has the full ranking, if she ever finds herself in a situation in which she has some set of acts available to her, then she can just choose the one act there that has the highest ranking.[36]

Notice that there is nothing in Savage that models this process. The agent doesn't have probabilities over which acts will be available to her, nor over processes that could lead her to believe that various acts will become available to her. Though Savage does successfully represent the preference an agent has over acts, it does not represent her reasoning about which acts might actually be available to her.

Since Jeffrey's theory *does* allow an agent to have credences over her own acts, we might think that this problem is solved. However, the omission remains. Indeed, there is nothing in Jeffrey's theory that models how an agent might identify her options. Options are simply *stipulated*, from the outside.[37] What an agent has control over is not derived from her theory of the world, as would fit with the naturalistic spirit that permeates the rest of Jeffrey's framework.

[36] Using some tie-breaking mechanism if need be.

[37] "From the outside" here could mean two things. It could mean that the modeller imposes the decision problem. Or, it could be that the options are somehow from the agent's perspective, but that the process of identifying options is not included in the model, but left vague. For example, Skyrms writes the following when describing Jeffrey's theory:

> In the application of the theory the decision maker can identify a partition of propositions that represent the alternative possible acts of her decision problem, and a partition representing alternative states of the world. (p. 505, 1994)

Since Jeffrey's theory is partition-invariant the latter bit isn't as important. However, identifying acts *is* important, and is not modelled in the framework. Nor are the ramifications of an agent identifying her options as such. This is the core issue present in Jeffrey's note on the kinematics of preference (1977). I discuss how my account helps Jeffrey overcome this issue in the third chapter. Indeed, a careful analysis of the ramifications of an agent identifying her options is the basis of the desirability tracking approach I provide in chapter 2.

When describing how acts work in Jeffrey's theory, Katie Steele and H. Orri Stefánsson write,

> In other words, the only thing that picks out acts as special is their substantive content—these are the propositions that the agent has the power to choose/make true in the given situation. (2020)

This, of course, fits with Jeffrey's statement that "an act is then a proposition which is within the agent's power to make true if he pleases" (1983, p, 84). In general, it is also the case that the set of acts is taken to "form a partition of the sure event" (p. 120, 1977).[38]

It is clear that act propositions are supposed to be propositions under the agent's control, but, so far, nothing from a Jeffrey agent's theory of the world tells us what she takes to be under her control. Used in this way, the theory is very much like Savage's: it provides a recipe for the agent to choose an act, should she find herself in a situation with options. For example, when describing how to use Jeffrey's theory to make choices, Arif Ahmed writes:

> More generally still: if $O_1, \ldots, O_n$ describe an agent's options on any occasion then it is rational to realize an option O_i if and only if it maximizes news value amongst the O_j; that is, if and only if $V(O_i) = \max_j V(O_j)$. (2014, p. 44)

Just as Savage's theory needs input (a set of possible acts) to be used as a

[38]But see *The Logic of Decision*, p. 84, in which Jeffrey also considers the possibility that an agent may abstain from making a decision, and thus we might think of their act as the sure event. In general, I do not consider this option here.

decision theory, so does Jeffrey's. *But it does not actually model the agent identifying what her options are.* So it fails naturalism.

If we want to use Jeffrey's theory to represent an agent as a *decision* maker, in the full-blooded naturalistic sense, then we need some way to identify decision problems. Or, more generally, we need some way to identify propositions over which the agent has control, *from her own perspective.* In other words, we want to *endogenize* control. Providing such an account is the aim of the next chapter.

Chapter 2

Endogenizing Control

"Knowledge enormous makes a God of me.
Names, deeds, grey legends, dire events, rebellions,
Majesties, sovran voices, agonies,
Creations and destroyings, all at once
Pour into the wide hollows of my brain,
And deify me, as if some blithe wine
Or bright elixir peerless I had drunk,
And so become immortal..."
— John Keats, *Hyperion*

"Deliberation–deciding what to do–is a matter not only of clarifying your preferences, but of identifying your options."
— Richard C. Jeffrey, 1977, "A Note of the Kinematics of Preference", *Erkenntnis*, p. 137

Chapter Summary

Decision theory, like Bayesian epistemology, is first-personal: we want to understand, from an agent's point of view, what the rational decision is. This involves using the agent's own beliefs (probabilities) and values (utilities) to determine the choice-worthiness of acts. Yet, decision problems themselves are given entirely *exogenously*; the acts available are stipulated to be so from the modeller's point of view. This is in tension with a first-personal naturalism. A more satisfactory account of the agent's point of view would have the decision context arise from the agent's *own* attitudes towards the world. In this paper, I show how we can extract the propositions over which an agent believes she has some degree of *control* from her attitudes. We recover the standard act-partition as a special case of more general deliberative contexts.

2.1 The Problem

When I deliberate, I do so over things which I consider to be in my control. This idea is old. Aristotle, in his *Nichomachean Ethics*, writes the following about things about which we don't deliberate: "The reason that we do not deliberate about these things is that none of them can be effected by our agency."[1]

Given that deliberation and control are so tightly intertwined, we might expect that theories of rational deliberation, and decision making more broadly, tell us

[1]From the *Nichomachean Ethics III, iii, 1-6, trans. H. Rackham, 1926.* Skyrms uses this passage and what comes before it to motivate causal decision theory (1984). Here I use it to emphasize that what matters for deliberation is *control*.

something about when an agent takes herself to be in control over something. As it stands, they are silent. As I described in Chapter 1, the decision theories currently on offer describe how rational agents make decisions when they have control over things, using as input the agents' values and their beliefs about how the world works. But this leave the objects of control to be entirely *exogenously* given.

Given the first-personal character of decision theory, it makes sense to desire an account of control that comes from the agent's own beliefs about how the world works. It would also be naturalistic. As I argued in the previous chapter, leaving control as an exogenous aspect of the agent's deliberation fails the richness condition of naturalism.

The goal of this chapter is to endogenize control. That is, I will show how we can extract from an agent's attitudes partitions over which it makes sense to say an agent views herself as having some degree of control.[2]

In addition to the naturalistic motivations, there are also decision theoretic reasons for wanting to endogenize control. We can understand the trajectory of some of the main decision theoretic frameworks on offer as one of increasing endogenization, in which certain aspects of the model are derived from an agent's attitudes as opposed to being given exogenously. The von Neumann Morgenstern framework showed us how to extract an agent's utilities over outcomes (1953). To do so, it relied on exogenously given probability distributions, gambles/acts, and control. Savage improved on the situation by endogenizing the probability distribution (1972). The objects of prefer-

[2]Really we will have a condition that indicates that an agent views herself as having control over a partition. Whether or not this condition is sufficient for control is an open question. I will discuss this in connection with reflection in §2.4.

ence were still given exogenously.[3] Jeffrey moved forward by endogenizing the objects of preference, which he did by restricting their closure to be merely logical instead of causal (1983). As stated above, this leaves control as an exogenously given factor. Here, I try to endogenize this extra piece.[4]

The plan is as follows. In § 2.2 I clarify how we should understand the approach I take by drawing a parallel between my approach and two other topics in philosophy: the principle of reflection, and Skyrms' pragmatic reduction of chance. In § 2.3 I describe my strategy for identifying control, which I call *desirability tracking*. I show how to use this basic idea to identify control in a series of more nuanced decision theoretic contexts. I formalize this basic idea in a series of successively more general definitions of conditions that may or not hold of an agent's attitudes. First I introduce a condition, **Strong Desirability Tracking**, that corresponds to an agent having complete control over a partition. Second I introduce a set of **Weak Desirability Tracking** conditions, that correspond to an agent having some degree of control across a partition. Finally, I introduce a condition, **Blackbox Desirability Promoting**, that corresponds to an agent being able to effect the world in a way that promotes good outcomes, but without necessarily being able to say anything more about how exactly this works. As I introduce these conditions I show how they are decision-theoretic duals to conditions of generalized learning.

[3]They involved *every* possible function from states to outcomes, even those that the agent thinks are impossible.

[4]Of course, there are other models of decision making that I left out here, such as the Fishburn model (1964) and the Luce-Kranz model (1974). Spohn provides a great overview of how these models connect to the Savage and Jeffrey models (1977). Spohn's paper is also important for our discussion of act probabilities, which is the focus of chapter 3.

2.2 Philosophical Cousins

2.2.1 Reflection

The approach I develop in §2.3 uses features of an agent's degrees of belief and desires in order to identify when it makes sense to say that she views herself as having control.

This strategy of using properties of an agent's attitudes in order to characterize when an agent takes herself to be in a certain context is shared by the Bayesian principle of *reflection*. One (informally stated) version of reflection is the following:

> An agent's current degree of belief in event A should be equal to her expected future degree of belief.[5]

As written this sounds like a normative constraint, since the word "should" is used. If we instead written the principle without "should", then we end up with a condition that can hold or not of an agent's degrees of belief:

> An agent's current degree of belief in event A *equals* her expected future degree of belief.

We can then use this condition to test whether or not an agent takes herself to be in a genuine learning situation. That is, if she takes herself to be in a learn-

[5]Huttegger identifies three different formal precisifications of reflection in Huttegger (2013). The informal version I write here is closest to his **R2**.

ing context in which she will respond rationally to the evidence. Huttegger puts the point nicely:

> The claim is that dynamic incoherence and violations of reflection are indicators of *epistemic irrationality*... Dynamic incoherence—understood in a tempered sense and applied to situations that fall within the scope of the theory of conditional expectations—as well as reflection are diagnostic of epistemic irrationality. The epistemic irrationality applies to how an agent updates beliefs since we have assumed that the agent is synchronically rational. Thus, as long as one ignores larger considerations, an agent cannot violate reflection and at the same time think that she will form her future degrees of belief in an epistemically rational way. If she does consider herself to be epistemically rational, then her probability measure should observe reflection. (p. 423, Huttegger (2013))

The last sentence expresses an important part of the structure: *if* an agent takes herself to be epistemically rational, *then* she will satisfy reflection. Note that this does not necessarily go the other way: she may satisfy reflection, and yet we/she may have reasons for thinking that she will not respond rationally to the learning context.[6]

One could view the account I provide in § 2.3 as sharing this structure: *if* an agent views herself as in a decision context (and she believes that she

[6]See Huttegger's discussion of this take on reflection and Dutch books for further details (section 3, 2013). His position builds on the work of a number of philosophers including Ramsey (1931); Skyrms (1990); Armendt (1993); Howson and Urbach (1993); and Christensen (1996).

is rational), *then* her attitudes will satisfy the conditions I identify. Just as with reflection, one might view this condition as *diagnostic* of rational deliberation. However, also just as with reflection, an agent may accidentally satisfy the condition, without actually viewing herself as rationally deliberating in a decision context. Whether or not this is the correct view in an open question. I discuss this more in §2.4.

2.2.2 The De Finetti-Skyrms' Reduction of Chance

My approach also has the feature that control only makes sense at the level of an agent's uncertainty about the world. This is in contrast to an account of control that would be based on properties of individual worlds themselves. What I mean by this is that what an agent has control over is not a property of any single possible world, but is a property of an agent's uncertainty over worlds.

This feature has an analogue in the de Finetti-Skyrms' pragmatic reduction of chance. Skyrms' writes the following about this approach:

> De Finetti is the kind of positivist who doesn't believe in chance—who regards the whole idea as metaphysical excess baggage—but still wants to give an account of the kind of Bayesian reasoning referred to in the last paragraph. He gives such an account by proving a famous *representation* theorem. In essence, this shows that *one who has degrees of belief which exhibit a certain symmetry behaves as if he believes in chances and is uncertain as to what the*

> *correct chance distribution is.* For de Finetti, this demonstrates that belief in the reality of chances is a difference that makes no difference; chances are, for him, simply an artifact of the representation theorem. (pp. 12-13, 1984)

De Finetti and Skyrms take chances to be (reducible to) features of our *uncertainty* about the world, instead of features of the world itself.[7] Thus their approach is different from a more *definitional* reduction of chance, in which chance is reduced to some believer-independent property of the world.[8]

My approach to agency shares this feature. An agent having control over something is not a property of any individual world. Instead, control emerges at the level of an agent's uncertainty. This is no accidental property of the account; it is essential for carrying out the project within the constrains of the naturalism sketched in chapter 1. All propositions are either true or false at any individual world, including propositions describing acts. Thus, if an agent is to view herself and her acts as part of the world, it cannot be that, at that world, she genuinely had a number of acts available to her.[9] Just as de Finetti's view regards chance as "metaphysical excess baggage", the naturalistic view I take here regards any kind of fundamental free choice as metaphysical excess baggage. In the language of chapter 1, such a view would violate *austerity*.

[7] Of course, since we are in the world, there is a *sense* in which chances are features of the world. But this is very different from the more standard way of thinking of chances are part of the world itself.

[8] Skyrms (1984) criticises such programs, for example, those of Van Fraassen (1977) and Kyburg (1978).

[9] Of course one *could* try to build some account where an agent *does* have a choice at a world, even if at that world she in fact determinedly does a single act. But such an account would be difficult to square with austerity.

2.3 Desirability Tracking

We are working in the framework Jeffrey provides in *The Logic of Decision*. The goal is to write down a formal condition that can tell us when an agent takes herself to have control over a partition. We think of this condition as a kind of test we can perform on an agent's attitudes.

The key insight I use is that deliberation and control are deeply connected. In particular, the condition I write down will exploit the fact that an agent deliberates over that which she takes herself to control. The connection with deliberation leads to a simple idea: if an agent views a partition as under her control, then, as she learns things that make different members of the partition more or less desirable, the changes to the probability that she assigns across the members of the partition should track the changes in the desirability. The probability across this partition tracks its desirability.

Consider the titan Hyperion, as he deliberates about whether or not to fight the young Apollo. Intuitively, Hyperion has control over this. Suppose that we wanted to be able to evaluate whether or not Hyperion views himself as having control over this partition, without just using our intuition. In particular, we would want to look at Hyperion's attitudes towards propositions, as captured by his probabilities and desirabilities. What kind of test could we perform? The key is to look at how Hyperion's credences shift as he learns things that make fighting more or less desirable.

Suppose that Hyperion is able to ask Phoebe, the titan goddess of prophecy, whether or not Apollo will fight alone, or with his sister Artemis. Suppose that

Hyperion believes that it is worth challenging Apollo if he fights alone, but not if Artemis lends Apollo her aid. That is, the desirability of the proposition "I will fight Apollo" is greater than the desirability of the proposition "I will not fight Apollo" if Apollo fights alone, but is less desirable if Artemis joins in the fray. Suppose also that Hyperion takes Phoebe's prophecy to be infallible.

We conduct our test by looking at Hyperion's conditional probabilities. Intuitively, if Hyperion has control over whether or not he fights Apollo, then, conditional on Hyperion learning that Artemis will help her brother, Hyperion's credence that he will attack Apollo goes down. This is because, if Artemis helps, then Hyperion would find attacking less desirable than abstaining, and, since Hyperion controls this partition, the probability of the less desirable proposition should decrease.[10] Similarly, conditional on Hyperion knowing that Artemis would not help her brother, Hyperion's credence that he will attack increases.

Making this intuition precise is the project of the present section. We will see that making things precise involves some subtlety. As we generalize this idea to more sophisticated decision contexts, we will be forced to write down more complex conditions. Despite these nuances, this desirability tracking behavior still forms the base of the test.[11]

The rest of this section proceeds by starting off with simple cases of control, and writing down corresponding desirability tracking conditions. As we get a

[10]Here I am relying on the idea that Hyperion responds *rationally* to this evidence, by using his control over the partition to increase the probability of the more desirable proposition. This is similar to using reflection as a test of rational learning, which I discuss in § 2.2.

[11]At least until the final case of *blackbox control* in §2.3.4, where we introduce a more general condition shared by all of the previous desirability tracking conditions.

grip on these we move to more complicated cases of control by relaxing certain assumptions. Ultimately, we will see that we can express the desirability tracking idea in a very general way that captures each of the previous versions as sub-cases.

2.3.1 Strong Desirability Tracking: Complete Control

First, we consider a case in which an agent has complete control over a partition: each member of the partition is a possible act for the agent. That is, the agent (intuitively) gets to make any single member of the partition true. This coheres well with Jeffrey's idea that an act is "a proposition which is within the agent's power to make true if he pleases" (p. 84, 1983) and is the standard case in decision theory. Throughout this analysis, we will suppose that agents view themselves as rational, in that they take correct actions given their beliefs and desires.

As a simple case, consider again our example from above: Hyperion is deliberating about whether or not to fight Apollo. Let $\{\text{Fight}, \neg\text{Fight}\}$ be the partition that corresponds to the possible outcomes of the deliberation. That is, Fight is the set of possible worlds in which Hyperion fights Apollo, and $\neg$Fight is the set of possible worlds where he does not.

Recall that our guide here is the intuitive idea that agents deliberate about things over which they have control. Thus, given that we are thinking of Hyperion as deliberating, this already tells us something about Hyperion's attitudes towards the partition: each member must have positive probability.[12]

[12]Of course, given that we are working in Jeffrey's framework, this is already a given. The

This is for the following reason. If Hyperion is already certain about one member of this partition (for example, if $P(\text{Fight}) = 1$), then we follow Skyrms in thinking that, for Hyperion, "there is no decision problem (i.e. his prior probability that he will choose a certain act is one.)" (p. 74, 1984). Thus, if we think that Hyperion can control this partition, that is, *if his deliberation will have any effect on which member is realized*, then Hyperion had better assign positive probability to each member of the partition.

Skyrms goes into more detail on the relationship between probability and deliberation:

> Indeed, one can argue that if a deliberator is absolutely sure which act he is going to do he needn't deliberate, and if he is absolutely sure he won't do one of a set of alternative acts his deliberations should concern only the others. Putting it the other way around, if a decisionmaker thinks that there is any chance that deliberation might change his probabilities of an act, he should have given the act a probability different from zero or one. (p. 36, 1990)

The account here is very much in agreement with this Skyrmsian position. However, we take an even more austere position. Whereas Skyrms allows agents to assign probability 0 to acts (and just not factor them into their deliberation), we do not countenance acts with 0 probability. I.e., whereas Skyrms might have an act in the set of possible acts that receives probability 0, on my account this would not even count as an act. This is driven by the

agent's beliefs are given by a strictly positive measure over the algebra. However, instead of merely reading this off the formalism, we will see that this actually has some philosophical justification.

austerity condition of naturalism described in the first chapter. If an agent is completely certain that something will not be the case, then it should not even be represented in the model.[13] Indeed, Jeffrey's framework enforces this, since agents' degrees of belief are given by a strictly positive probability measure.

This establishes that, if Hyperion views the {Fight, ¬Fight} partition as under his control, each member must get positive probability. This is the first, albeit quite weak, constraint on the partition.

More substantial constraints follow from Hyperion viewing the probabilities across this partition as driven by his deliberation. Since the agent is uncertain about which member of the partition is true, and since he views it as in his power to make true as he pleases, then there must be something which he expects he might learn that would change his preference ordering among the acts, before the time of decision.

For example, if Hyperion believes that he must act immediately, without any further deliberation, then he would choose the currently most desirable act. Suppose, for example, that this is the decision to fight. Then $P(\text{Fight}) = 1$, and, there is no decision problem.[14] Thus, if Hyperion does *not* have probability 1 in either proposition, this must mean it is because he thinks it is possible that he will learn something decision-relevant before he must act.[15]

[13]Though this position falls out quite naturally from from a naturalistic perspective, this point is actually contentious. For example, James Joyce argues against the position that epistemic impossibility implies pragmatic impossibility in a paper responding to the "deliberation crowds out prediction" thesis (2002). A very detailed discussion of how the present account of control bears on the act probabilities debate forms the bulk of the next chapter.

[14]Compare this with Jeffrey's discussion of the agent assigning probability 1 to not bringing wine in his discussion of acts on page 85 of *The Logic of Decision* (1983).

[15]Or, that he has only partial control over the partition. But in this section we are not concerned with that case. I discuss it in the following sections.

This process of (possibly) learning decision relevant things *is* deliberation.[16]

The *decision-relevant* property of the possible learning situation is essential: if Hyperion knew that, no matter what he would learn, he would still prefer fighting to not fighting, then he would already have probability 1 in fighting. Thus, as before, there would be no decision problem. Thus, if Hyperion is uncertain, then it must be because he assigns positive probability to an event where he learns something that flips his preference ranking, making not fighting more desirable than fighting.

Perhaps, as in the above example, Hyperion believes that he may be able to ask Phoebe whether or not Artemis would join Apollo, and assigns positive probability to both answers. Also, as before, Hyperion prefers not fighting to fighting if Artemis joins the battle. Since Hyperion believes himself to be rational, his probability of not fighting will increase if he learns that Artemis will join. In fact, since in this example we are supposing that Hyperion has total control over the partition, and we suppose that this is the only possible learning experience Hyperion expects to have before making his decision, the probability goes to 1.

In contrast, suppose that (intuitively) Hyperion did *not* view himself as having control over the partition. Then, even if he were to learn something that makes fighting more desirabile (Artemis will not join the battle), this would not change his credence that fighting will occur. Hyperion learning something that makes fighting desirable does *not* make it likelier.

[16] We often think of deliberation as something more internal to the agent. Many of the examples here use external sources of information. The account is compatible with any source of decision-relevant information. Thus, if Hyperion is uncertain because he is performing decision-relevant computations in his head that have not yet resolved, then this counts as deliberation.

Putting all of these ideas together, we can write down our first version of a desirability tracking principle. First, we need one extra bit of formalism that captures how we think of *learning*.

In order to make precise the different possible learning events the agent believes she might face, I use formal machinery derived from Matthias Hild's paper *Auto-Epistemology and Updating* (1998). I adapt it to the Jeffrey context, mainly by enforcing a measurability requirement to ensure that everything stays within the conceptual space of the agent. For those interested in the details: let I be a temporal index and let $\mathcal{EV}$ be the set of possible pieces of total evidence (for example, propositions or Jeffrey constraints). An *evidence function* $\pi : \Omega \times I \to \mathcal{EV}$ is a measurable function that assigns to each time i and each world ω the piece of total evidence that the agent receives. The event $L_i(e), e \in \mathcal{EV}$ is the event that the agent receives e as her total evidence at time i.[17] If I' is a sub-interval of I, then we write $\{L_i(e_i)\}_{i \in I'}$ for a sequence of learning events. In all of the desirability tracking conditions to come, $\mathcal{EV} = \mathfrak{A}'$ (recall that $\mathfrak{A}'$ is the domain of the agent's preference relation). I also assume that learning is veridical: $\omega \in \pi(i, \omega), \forall \omega \in \Omega, \forall i \in I$. The core thing to understand is that the possible pieces of evidence E_n are propositions that the agent might learn, and $L(E_n)$ is the event that the agent in fact learns E_n. Notice that, in general, $E_n \neq L(E_n)$. It is not necessary that, whenever E_n is true, the agent will learn that it is true. To illustrate, in the example above, if Hyperion thinks that Phoebe might ignore his question, then these two will come apart.

With this machinery on the table we can state our first Desirability Tracking

[17] I often drop the i when it is clear from context.

condition:

Strong Desirability Tracking. A partition $\mathbb{A} \subseteq \mathfrak{A}$ is *strongly desirability tracking* if for each member $A_n \in \mathbb{A}$ there exists a learning event $L(E_n)$, possibly $E_n = \top$, such that

1. $U(A_n|E_n) > U(A_m|E_n), \forall m \neq n$; and
2. $P(A_n|L(E_n)) = 1$.

Furthermore, there must be no learning event $L(E)$ such that (1) holds for some A_n, but (2) does not hold for that same A_n.

Let us walk through this condition. We will see that it captures all of the reasoning we did above.

First, it is a property that holds of a partition. This partition must be contained in the agent's algebra. The condition states that for each member of the partition there must be some corresponding learning event such that two properties must hold. The first condition states that the evidence that the agent learns in that learning event makes the corresponding member of the partition the most desirable of all the members.[18] This captures the idea of *decision-relevant* learning: each event in the partition has some piece of possible evidence that would make it the most desirable. The second condition states that, if the agent learns that piece of evidence, then the probability across the partition tracks this change in desirability. Indeed, since this condition captures a situation of full control, if the agent learns something that

[18]Here I use $U(A|E)$ to express the conditional desirability: $U(A|E) = \frac{1}{P(A|E)} \int_E u \mathrm{d}P_E = U(A \cap E)$.

flips the preference ordering,[19] then the probability of the most preferred act goes to 1.

Second, really, we should replace the single learning events in the Strong Desirability Tracking with possible sequences of learning events. For ease of exposition I write the conditions with single learning events, but it is easy to see how to extend them to sequences of learning events.

Notice also that I used a *strict* inequality in condition (1). This is slightly nonstandard, since we usually allow decision contexts in which there are two or more choices that maximize desirability. Thus, we might think to use a non-strict inequality. The reason I opt for the strict inequality here is that my analysis of control expresses the idea that it makes sense to say we control something if (its probability) can change (in the right direction) due to our rational deliberation. If two propositions are tied at the top of the preference ranking, then our *rational* deliberation will not influence which one gets realized. Normally we say something like, the agent uses a tie-breaking device to make her choice. Here, I would instead *externalize* that choice to the world, since the agent's deliberation does not affect what happens.[20] Instead, a different partition formed by taking the disjunction of tied acts (might) better express the agent's decision context (if this new partition also satisfies the desirability tracking condition).

One tricky thing is the remark that E_n will be the unit $\top$ for one of the A_n. The best way to see why this is needed is by thinking through the example.

[19] And she does not think it possible that she will learn something else decision-relevant before the time of decision.

[20] This corresponds to externalizing the agent's *type*, or how she *picks*, in the terminology of Joyce (2020). This makes sense, since how the agent picks is *not* something that is meant to be under her control.

First we will consider the learning events. Suppose that there are three possible things that Phoebe may do: she may honestly tell Hyperion that Artemis will join, she may honestly tell Hyperion that Artemis will not join, and she may ignore the question. Since Hyperion views them as possible, and since we are in Jeffrey's framework, Hyperion has positive degree of belief in each learning event.

Now, let us consider the two propositions in our control partition, and find their corresponding piece of evidence that Strong Desirability Tracking requires. The piece of evidence that corresponds to ¬Fight is the proposition "Artemis will join". This satisfies the first condition:

$$U(\neg\text{Fight}|\text{Artemis will join}) > U(\text{Fight}|\text{Artemis will join}).$$

Furthermore, $P(\neg\text{Fight}|L(\text{Artemis will join})) = 1$, since Hyperion is rational, and has control over the partition.

So far so good. But consider the other proposition in the partition, Fight. We might think that the corresponding piece of evidence is Artemis will not join. However, this runs into a difficulty. The condition

$$U(\text{Fight}|\text{Artemis will not join}) > U(\neg\text{Fight}|\text{Artemis will not join})$$

cannot be met, since there is no world in which Artemis does not join the fight and Hyperion doesn't fight: thus the desirability of Hyperion not fighting and Artemis not joining is undefined, as their conjunction is the zero F. This is because fighting is Hyperion's *default* action: if he doesn't learn anything (Phoebe ignores him), then he fights anyways. Thus, if Artemis doesn't join,

then Hyperion either doesn't learn this, and fights, or he does learn this, and fights. Either way, Hyperion fights.

However, we see that there *is* some piece of evidence that satisfies the two conditions: $\top$! The desirability of Hyperion not fighting is perfectly well defined given $\top$, and is less than that of fighting, since fighting is the originally preferred member of the partition. And, furthermore, $P(\text{Fight}|L(\top)) = 1$, since it is preferred under the agent's prior. Indeed, having $\top$ as the corresponding piece of evidence for Strong Desirability tracking witnesses the fact that fighting is Hyperion's default act.

Thus, we see that Strong Desirability Tracking is a test we can perform on a partition. If the partition fails this test, then we can conclude that the agent does not view herself as having control over this partition. When a partition satisfies this condition, this provides a way for an agent to view herself as making decisions, while satisfies the constraints of naturalism.

2.3.2 Weak Desirability Tracking: Trying

In § 2.3.1 we saw how to extract an act partition from an agent's degrees of beliefs and desires. This captures the core case studied in decision theory, in which agents are able to simply make a proposition true.

There are other, more general cases in which we might be interested. These cases mirror the generalization of learning from conditioning on a member of a partition to merely shifting the probability across a partition. Learning a member E_j of a partition $\mathbb{E} = \{E_i\}_{i \leq n}$ for certain involves moving from a

probability distribution

$$P(E_1), P(E_2), \ldots, P(E_n)$$

to a distribution where $P(E_j) = 1$ and $P(E_i) = 0, \forall i \neq j$. In contrast, a more general form of learning might have

$$P(E_1), P(E_2), \ldots, P(E_n)$$

change to some other distribution

$$P'(E_1), P'(E_2), \ldots, P'(E_n)$$

without necessarily having the probability of any single member go to 1. This, introduced by Jeffrey, is called *probability kinematics* (1983). Learning by probability kinematics, and indeed more general forms of learning, have been studied extensively.[21]

Jeffrey identified *probabilistic acts* as the decision theoretic analogues to probability kinematics:

> The situation that we have been studying in relation to probabilistic observations has its parallel in the case of probabilistic acts. It may be that the agent decides to perform an act which is not simply describable as *making the proposition B true*, but must be described as changing the probabilities of two or more proposi-

[21] See, for example, Jeffrey (1983); Levi (1967); Van Fraassen (1980); Skyrms (1987); Skyrms (1990); and Huttegger (2015).

tions... from

$$prob\ B_1, prob\ B_2, \ldots, prob\ B_n$$

to a new set of values,

$$PROB\ B_1, PROB\ B_2, \ldots, PROB\ B_n.$$

(p. 177, 1983)

Furthermore, he identified a particular type of probabilistic act he called *trying*:

> In the simplest cases, where $n = 2$, where B_1 is some good proposition B and where B_2 is the bad proposition $\overline{B}$, we speak of the agent as *trying to make B true*; and where $PROB\ B$, the probability that B would have if the agent decided to perform the act, is very close to 1, we may speak of the agent as believing it to be in his power to make B happen if he chooses. (p. 177, 1983)

Just as probability kinematics generalizes learning from conditioning on a member of a partition to shifting probability across a partition, probabilistic acts and trying generalize having complete control to having a weaker form of control.

We want to know if we can endogenize this weaker form of control. That is, can we identify a condition on the agent's attitudes that will be satisfied when an agent views herself as having this weaker form of control. I will show that

we can do this. The key insight again comes from appropriately formalizing desirability tracking.

First, in this section I will show how we can endogenize the simple case of binary trying. This will introduce some new nuances into our thoughts around desirability tracking.

Consider once again the case of Hyperion deliberating about whether or not to fight. Instead of him having complete control, however, imagine that he can only partially influence whether or not he will fight. In Jeffrey's words, he can *try* to fight or not, but he doesn't expect himself to be able to perfectly control whether or not he does.

Once again, if there is nothing that Hyperion could learn that would flip his preference between fighting and not fighting, then he faces no decision problem. Thus our analysis will again include some possibility of learning. In our running example, suppose that he is still able to ask Phoebe whether or not Artemis will join the fight. This is again our possible learning event. Assume that Hyperion's preferences are the same as before.

Then, instead of the probability of some member of the $\{\text{Fight}, \neg\text{Fight}\}$ partition going to 1 based on what Hyperion learns, in should merely shift in the right direction. This leads us to our next type of desirability tracking:

Weak Binary Desirability Tracking. A binary partition $\{A, \neg A\} \subseteq \mathfrak{A}$ is *weakly desirability tracking* if for each member $A_n \in \{A, \neg A\}$ there exists a learning event $L(E_n)$, possibly $E_n = \top$, such that

1. $U(A_n|E_n) > U(\neg A_n|E_n)$; and

2. $P(A_n|L(E_n)) > P(A_n|E_n)$.

Furthermore, there must be no learning event $L(E)$ such that (1) holds for some $X \in \{A, \neg A\}$, but (2) does not hold for that same X.

Notice that Weak Binary Desirability Tracking differs from Strong Desirability tracking in ways other than the restriction to binary partitions. First, we do not require that the probability of the preferred proposition go all the way to 1. This reflects the fact that the agent can only *try* to make A true.

Second, we capture the idea that the probability of the preferred act moves in the correct direction by requiring that its probability conditional on learning the corresponding evidence be greater than merely its probability conditional on the evidence. We require this for two reasons. First, this allows for more flexibility: instead of a fixed way in which the agent can "nudge" the probability of events, we allow the "default" probabilities to depend on which piece of evidence is learned.[22] For example, it may be that Hyperion can nudge the probability of fighting more if Artemis joins the fray than if she doesn't. Second, this allows us to determine that the agent learning the evidence (and having the corresponding shift in probability) changes the probability in the correct direction, more than just the evidence itself. This captures the idea that the agent is efficacious.[23]

Here a subtlety accosts us. The issue is one of *rigidity*. In learning by probability kinematics, not only does the probability across a partition shift, but

[22]What this means is that the "default" probabilities here are given not by $P(A), P(\neg A)$, but by $P(A|E), P(\neg A|E)$.

[23]Basically, learning $L(E_n)$ induces a Jeffrey shift across the $\{A, \neg A\}$ partition. The careful reader will wonder about why/how we think of this as a Jeffrey shift. The following discussion of rigidity addresses this.

it shifts in a way that leaves conditional probabilities unchanged.[24] Skyrms gives us a way to think about rigidity in terms of a sufficiency condition:

> Jeffrey uses the name "probability kinematics" to suggest the absence of information forces that might deform the probabilities conditional on members of the partition. In statistician's language the partition $\{e_j\}$ is a *sufficient partition* for the class of probability distributions that can come from Pr_i by probability kinematics on $\{e_j\}$, and a measurable function whose set of inverse images is the partition is a *sufficient statistic* for that class of probability distributions. (p. 6, 1987)

This *sufficiency* condition ensures that the change to the probabilities of different events in the algebra is entirely filtered through the change of the probabilities across the evidence partition. This allows the shift to propagate its effects through the rest of the algebra in a principled way.

A similar condition should hold for the analogous probabilistic acts; the effect of the agent trying to make some proposition true should be filtered entirely through the change in the probability of the partition. Stating this condition in the context of probabilistic acts is trickier, and new territory. Jeffrey doesn't discuss this issue at all in *The Logic of Decision* (1983).

We might think that we can enforce rigidity in exactly the same way. We might try to require that, for each E_n that the agent can learn, and $\forall A_m \in \{A_i\}, \forall B \in \mathfrak{A}, P(B|A_m \wedge E_n) = P(B|A_m \wedge L(E_n))$. The intuition is that all of

[24] That is, $\forall A \in \mathfrak{A}, P(A|E) = P'(A|E)$ where P is the probability function before the shift and P' is the probability function after the shift.

the effect of learning E_n is via the shift across the partition $\{A_i\}$. This cannot in general work, however. Any proposition in $A_m \cap E_n$ that is dependent on $L(E_n)$ will make violate this condition, including $L(E_n)$ itself. So, the simple approach will not work.

We have at least three options we might consider. The first is to try to enforce some kind of restricted version of rigidity analogous to the rigidity in Jeffrey shifts. The second and third options both to a more desirability oriented approach, where instead of enforcing rigidity on probability we force it on desirability, in different ways.

The first option is similar to what Hild does when he states an auto-epistemic version of rigidity, which he calls *Evidential Independence*. Hild's constructions relies on a base algebra $\mathcal{A}^0$, that is then enriched via auto-epistemic vocabulary. Thus, using our notation, Hild's Evidential Independence condition states that for all members A_n of the evidence partition $\mathbb{A}$, and $\forall B \in \mathcal{A}^0$,

$$P(B|A_n \wedge L(E_i)) = P(B|A_n).$$

This says that rigidity holds with respect to all members of the base algebra, but not necessarily for the enriched algebra. We might take a similar route, separating out all of the propositions about the epistemic state of the agent, and requiring rigidity to hold with respect to those.

The second option involves enforcing a kind of *value rigidity*. In our context, this would be the condition that $U(A|E) = U(A|L(E)), \forall A \in \mathbb{A}, \forall E_A : A \in \mathbb{A}$.[25] In words, the desirability of each member of the trying partition depends

[25]Recall that, since learning is factive, $L(E) \subseteq E$, and so this condition is also equivalent

only on that member and which piece of (decision relevant) evidence is true, but not on the fact that the agent has learned that evidence. This would help to ensure that the change in value for the agent of $\top$ before an after learning E is entirely filtered through the change across the partition $\mathbb{A}$. This is the sense is which this is a desirability version of rigidity.

A condition like this is often assumed in decision theoretic contexts. For example, this is the second half of the second condition that Skyrms' identifies holds when proving the value of information theorem: "Secondly, by using the same notation for acts and states pre- and postexperiment, we are assuming that performance of the experiment itself cannot affect the state in any relevant way and that, so far as they affect consequences, the generic acts available postexperiment are equivalent to those available preexperiment" (p. 246, 1990). Here Skyrms is operating within a Savage-style framework, but the core idea is the same: the learning itself doesn't affect the expected utility of the acts, other than through how it provides information to the agent *for that decision*.

We often understand this condition as saying that learning is *cost-free* (chapter 4, Skyrms (1990)); the learning experience itself doesn't impose any costs on the agent. It actually goes the other way as well—the learning experience itself doesn't provide any benefits to the agent, other than how it changes her disposition to act in this particular problem. That is, the learning experience is both cost-free and benefit-free, modulo the action of the agent in the particular decision context at hand.

Under this second option, the intuitive story goes as follows. The agent learns

to $U(A|E) = U(A|E \wedge L(E)), \forall A \in \mathbb{A}, \forall E_A : A \in \mathbb{A}$.

E, which makes A the most desirable proposition. Then, she exerts her influence on the world, which only makes the world more desirable for her insofar as A becomes more likely; the value of A itself, and of $\neg A$, does not change.

These first two options have the virtue of capturing the surgical notion of a probabilistic act. The first one does so by specifying that the probability of events in some basic algebra are rigid, and the second one does so by specifying that the value within each member of the partition remains unchanged by conditioning on the learning event.

By the same token, they also have the vice that they rule out the agent benefiting in the future due to a gain of information. The first option does so by stipulating that the events in the base algebra are independent of the event that the agent has learned something. But this rules out, from her perspective, the possibility that this information will prove useful to her in some future decision.[26] The second option does this by enforcing that the value within a proposition formed by intersecting a member of $\mathbb{A}$ and an event the agent can learn doesn't change based on whether or not the agent has learned the evidence. It might be the case that the evidence helps the agent in future deliberations, but, if so, this is perfectly countered-balanced by some cost induced by a correlation between the learning event and some undesirable state of affairs.[27]

This kind of condition makes sense when the evidence the agent gains for deliberation is only useful for the task at hand, and when, from the agent's

[26]Otherwise, there would be some member of the base algebra that would *not* be independent of the event that the agent learned something, that made it more/less desirable than another proposition.

[27]It might also be the case that the evidence doesn't help the agent in future decisions.

perspective, the gathering of the information is cost-free in the sense that it does not affect the state of the world in some bad way, other than through what it tells the agent.[28] Thus, insofar as we want to capture this kind of fairly local deliberation, these options are good for the job.[29]

However, if we want to allow the same information might be relevant in future deliberation, enforcing this kind of rigidity will not do. This brings us to the third option: we might require that $U(A_n|E_n) \leq U(A_n|L(E_n)), \forall E_n, A_n : A_n \in \mathbb{A}$. This would ensure that the learning itself is cost-free, but allows for a potential benefit beyond how it informs the agent's current decision.

There is also of course a fourth option: don't require anything, and stick with Weak Desirability tracking as written. How do we think of this?

At this point it is helpful to recall the kind of project I am carrying out. I am identifying conditions such that, if an agent's attitudes towards certain event satisfy them, it indicates that an agent views herself as having control. With this spirit, if we notice that, for an agent with certain beliefs and desires, a binary partition satisfies Weak Binary Desirability tracking, with some specified version of rigidity, then we can say that the agent views herself as performing the kind of probabilistic trying that Jeffrey described. If, on the other hand, we notice that the partition satisfies Weak Binary Desirability Tracking, and the partition fails the rigidity requirement, then we can infer that either she views her deliberation as not merely concerned with that partition, or there is some kind of correlation between her learning E and other propositions of

[28] Once again, see the second condition needed for the value of information theorem on page 246 of Skyrms (1990).

[29] Local in the sense that the information generated will not help the agent in future decision contexts.

interest that is not filtered through her deliberation.

With this perspective, it is a mistake to try to say *which* condition *should* be satisfied. Rather, we say that *if* such a condition *is* satisfied, then we can make the corresponding statement about how the agent views her decision context. With this in mind, I state the following refinement of weak binary desirability tracking, which gives us a more surgical version of *trying*:

Surgical Weak Binary Desirability Tracking. A binary partition $\{A, \neg A\} \subseteq \mathfrak{A}$ is *surgically weakly desirability tracking* if for each member $A_n \in \mathbb{A}$ there exists a learning event $L(E_n)$, possibly $E_n = \top$, such that

1. $U(A_n|E_n) > U(\neg A_n|E_n)$;
2. $P(A_n|L(E_n)) > P(A_n|E_n)$; and
3. $U(A_n|E_n) = U(A_n|L(E_n))$.

Furthermore, there must be no learning event $L(E)$ such that (1) and (2) hold for some $X \in \{A, \neg A\}$, but (2) does not hold for that same X.

Surgical weak desirability tracking includes the value rigidity condition. Thus it captures a case in which the only way the agent makes the world better for herself, on net, is through changing the probability of A.

2.3.3 Weak Desirability Tracking: Probabilistic Acts

Now we consider the situation in which the number of members in our partition is $n \geq 2$, which extends the binary case. Here, the intuition is exactly the same

as in the binary trying case. All we need is one additional technical condition, which we got for free in the binary case.

In the binary case, whenever the probability of A increased, the probability of $\neg A$ necessarily decreased, since $P(\neg A) = 1 - P(A)$. This meant that if the agent increased the more desirable outcome then, overall, things became better for her in expectation (modulo possible worries about cases in which value rigidity is not satisfied).

This doesn't necessary hold when $n \geq 3$. Consider, for example, a partition with three members: A_1, A_2, and A_3. Suppose that, conditional on E_1, the agent's preference ranking is $A_1 \succ A_2 \succ A_2$. If the probability across the partition shifts such that $P(A_1|E_1) > P(A_1|L(E_1))$, this does not ensure that overall the probabilistic act will improve things. For example, if A_3 is greatly dispreferred to A_2, and in the new probability distribution A_3 has increased its probability by "stealing" probability from A_2, then even though the probability of the most desirable proposition has increased, overall things have gotten worse for an agent. Given that things should not get worse for an agent if she is rational and exerting control, we want to rule these cases out.

This leads us to our next desirability tracking condition:

Weak Desirability Tracking. A partition $\mathbb{A} = \{A_n\} \subseteq \mathfrak{A}$ is *weakly desirability tracking* if for each member $A_n \in \mathbb{A}$ there exists a learning event $L(E_n)$, possibly $E_n = \top$, such that

1. $U(A_n|E_n) > U(A_m|E_n), \forall m \neq n$;
2. $P(A_n|L(E_n)) > P(A_n|E_n)$; and

3. $P(A_m|L(E_n)) \leq P(A_m|E_n), \forall m \neq n.$

Furthermore, there must be no learning event $L(E)$ such that (1) holds for some $X \in \{A_n\}$, but (2) does not hold for that same X. Furthermore, if $U(A_n|E_n) = E(A_n|L(E_n))$ for all A_n, then we say that $\mathbb{A}$ is *surgically* weakly desirability tracking.

If a partition satisfies this condition for an agent, it indicates to us that she views herself as having probabilistic control over the partition, in the sense that she is able to increase the probability of the most desirable member of the partition.

2.3.4 Blackbox Desirability Promoting

So far there is a parallel between more and more generalized learning, and more and more generalized notions of control. At the extreme end of the former we have *blackbox learning*, first studied by Skyrms (1990). Blackbox learning describes a situation in which an agent undergoes some kind of learning experience about which she can't articulate the details, except what the possible belief-outcomes of the experience are. More carefully, blackbox learning occurs when your current beliefs are captured by some probability function P, and there is some set $\{P_i'\}$ such that you expect that, after some (possibly) unspecified learning experience, it is possible that any of the $P' \in \{P_i'\}$ will become your degrees of belief. This is *blackbox* because it is not (in general) the case that the belief change comes about in some cleanly specifiable way, such as in probability kinematics.

Despite its opaque nature, we can still identify a necessary condition for an agent to view a blackbox belief-change event as as learning event. This is (no surprise) the principle of reflection:[30]

$$P(A|P_i') = P_i'(A).^{31}.$$

We want to know: can we identify a necessary condition for blackbox control? In each of the previous cases we considered we could say something principled about how the agent learning some decision-relevant proposition E would change the probabilities in the desirable direction. In each case there was a partition, and the agent's control consisted in shifting the probabilities across the partition so as to make the most desirable state more likely. Blackbox control throws all of that structure out. We imagine that the agent views herself as making herself better off, but that she (might) be unable to articulate exactly *how* she make things better off for herself. This is analogous to blackbox learning; the agent can't necessarily say what she learned or why she changed her beliefs in the way she did, only that she takes the experience to be one of rational learning.

This generalized idea of control leads us to the following condition, which holds of a learning experience. Here, a learning experience is a measurable function π from possible worlds to evidence. For simplicity, I write the following condition where π maps to propositions that the agent learns.

Blackbox Desirability Promoting. A learning experience π is *desirability*

[30] See Huttegger (2014) for a detailed discussion of how reflection relates to decision making.

[31] Supplemented with appropriate quantifiers.

promoting if for each possible learning event $L(E)$[32] the following condition holds:

$$U(L(E)) > U(E).$$

A learning experience is desirability promoting if the agent thinks that situations where she learns the evidence have a higher expected value than situations when the evidence is true. It turns out that all of the other desirability tracking conditions above satisfy this, but in more structured ways. In general, the expected desirability of learning E is higher than that of just E because the agent pushes the probability of the most desirable member of a partition, given E, in the right direction. In the case of blackbox control we lose this detailed structure but keep the reward: however she does it, the agent makes things better off (in expectation).

2.4 Necessary or Sufficient

For the case of Blackbox Desirability Promoting, it seems like this condition doesn't imply that an agent takes herself to have some kind of control. She could, for example, just intrinsically value knowing true things, and so prefers worlds in which she learns E to those in which she doesn't. However, if she does take herself to be able to make things better off for herself as a result of deliberation, then this condition will hold.

There is an interesting question here about the conditions under which we can

[32]Where a possible learning event of π is a proposition $L(E)$ such that E is in the range of π, i.e., E is a proposition the agent can learn as a result of π.

separate the intrinsic value an agent might assign to knowing things from the more instrumental value that learning has. However, this also refocuses us on a previous question to which I alluded in § 2.2.2: are the various desirability tracking conditions merely necessary, or actually sufficient, for control?

I take this to be an important open question. I'll briefly describe two ways in which one might supplement the desirability tracking account that some might find attractive. Ultimately, however, in chapter 3 I will work with a desirability tracking account of control as if it were sufficient. This will show that, even if ultimately one did want to further supplement the account, it still can do a lot of philosophical work.

2.4.1 Ramsey Thesis

The *Ramsey Thesis* offers one further refinement of a desirability tracking account of control. The idea of the Ramsey Thesis comes from a passage of Ramsey's 1929:

> What is true is this, that any possible present volition of ours is (for us) irrelevant to any past event. To another (or to ourselves in the future) it can serve as a sign of the past, but to us now what we do affects only the probability of the future. (p. 158)

This passage is often interpreted as putting a restraint on the kinds of attitudes that a deliberating agent can have. For example, Arif Ahmed summarizes a consequence of the Ramsey Thesis (what he calls *Evidential Dualism*) as follows:

> Evidential Dualism is a claim about what you should *think*. It says that a rational agent contemplating an act will not take it to stand in the same relation of evidential relevance that she would if viewing it as an observer. (p. 217, 2014)

The core idea is that, whenever an agent is deliberating about what to do, her beliefs should satisfy certain properties. For our present purpose, we might flip this. Instead of asserting the normative constraint that an agent must view herself in a certain way when deliberating, we might view the constraint as given a refinement of our desirability tracking account. This actually fits well with something Ramsey writes in the same section:

> This seems to me the root of the matter; that I cannot affect the past, is a way of saying something quite clearly true *about my degrees of belief.* (p. 158, 1929, emphasis added)

Thus, we might add to the various desirability tracking accounts the further condition that an agent's degrees of belief over the elements of the partition are independent of anything in the past.

I am not, however, sanguine about this approach. It seems far too restrictive. Ahmed gives a compelling series of arguments against the more normative flavour of the Ramsey Thesis, which extend also to the more diagnostic flavour expressed here (chapter 8, 2014).

To me, the Ramsey Thesis is best understood as an early attempt to put some kind of causal thinking into decision making. Indeed, Ramsey himself writes

that, when decision making, we are engaged with "tracing the different consequences of our possible actions, which we naturally do in sequence forward in time, proceeding from cause to effect not from effect to cause" (p. 158, 1929).

This suggests that we might try to capture the causal intuition more directly, which motivates the next possible refinement.

2.4.2 Causal Probability

All of the desirability tracking conditions we considered used conditional probability in their statements. This captured the idea that, upon learning something that was relevant to the ranking of the members of the partition, the probability would move in the right direction. A natural modification of this is to replace the conditional probability, $P(\cdot|\cdot)$, which captures evidential relationships, with a *causal* probability, often denoted $P(\cdot\backslash\cdot)$. $P(\cdot\backslash\cdot)$ is meant to capture the causal influence of the second argument on the first. Usually this is used to make causal decision theory precise, by taking the *causal expected utility* to be of the form

$$CEU(A) = \sum_S P(S\backslash A)u(A\&S).$$

There are many different ways to make the causal probability formally precise.[33] Given the general background Jeffrey framework I favour for naturalistic reasons, I suspect that going with an imaging-based formulation of the causal probability function, which Joyce uses to make a partition-invariant

[33]Joyce (1999) gives a thorough survey of various approaches in pages 161-180.

version of causal decision theory in Jeffrey's general framework, is the most attractive option. If one takes this approach, an interesting further question is whether the desirability parts of the desirability tracking account should *also* be replaced with causal versions. I leave this for future work.

2.5 Conclusion

We have identified a series of more general conditions that indicate that an agent takes herself to have some degree of rational control. These conditions reflect the condition of reflection for rational learning. They also share a core feature of the de Finetti-Skyrms reduction of chance: control, like chance, emerges out of an agent's uncertainty about the world, instead of within the world itself.

Chapter 3

Between Two Camps: A Ridge with a Branch

"More thought than woe was in her dusky face,
For she was prophesying of her glory;
And in her wide imagination stood
Palm-shaded temples, and high rival fanes,
By Oxus or in Ganges sacred isles."
— John Keats, *Hyperion*

Chapter Summary

Can an agent assign probabilities to her own acts while she deliberates? This question splits decision theorists into two broad camps: those that believe deliberation crowds out prediction, and those who believe that deliberation

welcomes prediction. I use the desirability tracking account of control developed in the previous chapter to chart a middle, naturalistic course. I argue that desirability tracking allows us to avoid the violations of *richness* that the crowding out camp commits, while also avoiding the violations of *austerity* that the welcoming camp commits.

3.1 A Ridge and Two Camps

Can an agent have beliefs about her own actions as she deliberates about what to do? In particular, can she assign them probabilities?

The answer may seem obvious. However, in good decision theoretic tradition,[1] to different people, *different answers* are obvious. For example, Wolfgang Spohn writes:

> Now, probably anyone will find it absurd to assume that someone has subjective probabilities for things which are under his control and which he can actualize as he pleases. (p. 115, 1977)

Contrast this with Alan Hájek's conclusion to his paper "Deliberation Welcomes Prediction":

[1] Robert Nozick writes this about the Newcomb Problem:

> To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly. (p. 117, 1969)

Though the act-probabilities debate has perhaps not caused quite as much ink to be spilled as the Newcomb problem, the ink-spillers have been just as passionate.

> Now that I'm done with this paper, I'll reward myself with a glass of wine. Now that you've finished reading it, you might want to do so too. I'm highly confident, but not certain, that I'll choose the red. How about you? (p. 526, 2016)

Hájek's response, read between the lines, seems to say: *of course* we assign probabilities to our own acts—we're doing it right now!

Luckily, we have more than intuitions to go on in our investigation. This question has spawned a rich literature, settling into two rough camps. One camp holds that *deliberation crowds out prediction*; the other that deliberation *welcomes* prediction. Both camps are full with the tents of philosophers. So why aren't all their tents pitched on the same hill?

In this chapter I pitch my own little tent on a ridge that lies between both camps. I invite everyone from both camps to join me on my ridge. It will appeal to those who take deliberation to welcome prediction by having a clear, unambiguous place for act probabilities. My ridge will also appeal to those who believe that deliberation crowds out prediction by showing that we can have act probabilities without appealing to powers and distinctions that go beyond the agent's theory of the world, and without robbing the agent of her agency.

The ridge also helps us understand why philosophers aren't all in the same camp: those in each camp have correctly identified deficiencies with the other camp. Without knowledge of the ridge, they have chosen the camp that seemed least deficient. My invitation is meant to lift the ridge out of the fog of war, and provide an adequate home for those of both camps. Indeed, I believe I

see, just outside my tent, the branch of an olive tree waving in the breeze.

Of course, not everyone will accept my invitation. Some, present in both camps, will think the notion of agency we have on my ridge is too weak. I will argue that how they choose to respond to this weakness is what drives them to seek one camp or another. One can acquire a stronger notion of agency by either leaving the agent outside her (decision-theoretic) model of the world (violating richness), or by adding basic powers and/or options that go beyond the agent's theory of the world (violating austerity). However, for those who want to have a naturalistic picture of agency, one that satisfies the richness and austerity conditions laid out in chapter 1, and are happy with a weaker notion of agency, my ridge offers just that.

In the rest of the chapter I chart you a path to my ridge. To orient ourselves, I will provide a survey of the arguments for the deliberation crowds out prediction thesis that I believe have been sufficiently addressed. First I will discuss arguments concerning probabilities as betting dispositions. I will summarize the core counter-arguments made against these arguments, which I find persuasive. Next, I will discuss Spohn's argument that there is no decision-theoretic role for act-probabilities, which I believe has some solid arguments against it, but for which desirability tracking provides an additional clear answer. After this I discuss what I take to be the strongest argument against act probabilities: Isaac Levi's argument from the vacuity of rational principles. I will argue that the main counterarguments in the pro-act-probabilities camp violate the austerity condition of naturalism. I will then describe how an understanding of deliberation that comes from desirability tracking does justice to the strong arguments against act probabilities, while still making room for

act probabilities. The ultimate upshot is that while we cannot save a moment of genuine choice, we can save deliberation. I will conclude by describing why one might still not be happy on my ridge, and how the current camps have avoided the ridge by violating the requirements of naturalism in different ways.

3.2 Probability as Betting Dispositions

One type of argument that the crowd-out camp deploys depends on the identification of probabilities with betting dispositions. Hájek has helpfully identified two different arguments that make use of betting rates (2016). These are what he calls *Betting rates collapse* and *Betting rates cannot be applied*, which he attributes to Levi and Spohn respectively. We will examine each in turn.

3.2.1 Betting Rates Collapse

Levi's argument relies on the identification of probabilities with betting dispositions (1991, 2000, 2007). This is related to the idea that the probabilities of a rational agent are revealed by an agent's betting behaviour. Historically this kind of idea goes back to Ramsey (1931) and de Finetti (1992).[2] There are stronger and weaker ways of understanding the connection between betting rates and degrees of belief. On a weak way, we think that, under some ideal conditions, probabilities are measurable by betting dispositions. On a stronger way, probabilities are identified with betting dispositions. Levi's argument relies on the strong version of this connection: the claim that an agent

[2]For a gentle introduction to this idea, see Chapter 2 of *Ten Great Ideas About Chance* (Diaconis and Skyrms (2017)).

only has a subjective probability for a proposition if she has a definite fair price for various bets involving that proposition.[3] Joyce, responding to Levi, states this premise as follows (2002):

> *Act Probabilities are Revealed in Fair Prices.* [The decision maker] has a definite subjective probability $P(A)$ for an act A if and only if $P(A)$ is her fair price for the wager $W_A = [\$1 \text{ if } A; \$0 \text{ if } \neg A]$ or, equivalently, she prefers $W_A(P(A))$ among all wagers of the form
>
> $$W_A(p) = [\$(1 - (1 - p)^2) \text{ if } A; \$(1 - p^2) \text{ if } \neg A]$$
>
> (2002, p. 82)

Supposing that one accepts this, we can see how this leads to Levi's desired conclusion. My discussion here follows Joyce's (2002) particularly clear reconstruction of Levi's arguments in Levi (1991) and Levi (2000).[4] Suppose that the agent is in a context where there are only two acts available to her, A and $\neg A$. Suppose furthermore that the agent strictly prefers A to $\neg A$, and that she is certain that she is rational.[5]

Suppose we attempt to measure her degree of belief in A before she has chosen to make A true or false.[6] If we do so using the strategy of offering her bets as described above, then her set of available options changes. Instead of choosing between A and $\neg A$, the set of acts she chooses between is the common

[3] This strong view has been deployed in other contexts. For example, in *The Emergent Multiverse*, David Wallace relies on such a view in order to recover probabilities in his decohering formulation of quantum mechanics (2012).

[4] As well as Hájek's reconstruction (2016).

[5] That is, she is certain that she will choose optimally given her beliefs and desires.

[6] Which she can do, as it is assumed that A and $\neg A$ are under her control.

refinement of the partitions $\{A, \neg A\}$ and $\{W_A(p)\}_{p\in[0,1]}$.

Then, Levi reasons, since she has control over $X \in \{A, \neg A\}$, the agent, being rational, would not make X true and then set $P(X) < 1$,[7] for she would be losing out on a sure gain of money. Thus, we can think of the agent as really choosing between $A \wedge W_A(1)$ and $\neg A \wedge W_A(0)$. But, since the agent gets a dollar either way, and since she is rational and knows she is rational, she will just choose the first option.[8] But then we see that this procedure reveals that her degree of belief in A is 1. But then this means she assigns a probability of 0 to $\neg A$, which was inadmissible for her (since she is rational and strictly prefers A to $\neg A$). And thus we have the desired conclusion.

This argument depends crucially on the *Act Probabilities are Revealed in Fair Prices* premise. The standard response from the other camp is that this premise is false. For example, when analysing the betting measurement procedure (about a different situation in which, in the absence of any bets, the agent prefers act B to act A), Wlodek Rabinowicz writes:

> If no bet on A is offered to the agent, the agent does not think it probable he will perform A. $P(A)$ is relatively low. But, if a bet on A is offered, with the net gain equal to G, $P(A)$ increases... Thus, *the probability of an action depends on whether the bet is offered or not.* But then the probability of an action cannot be measured by the agent's betting rate, for the offer of a bet itself changes the agent's probabilities. (p. 101, 2002)

[7] That is, set W_A's price at a value less than a dollar if $X = A$ and more than zero dollars if $X = \neg A$.

[8] Because we have assumed that she prefers A to $\neg A$.

Rabinowicz is pointing out that, when an agent has control over some proposition, then measuring an agent's credences by offering them bets is a *disturbing* measurement: it changes the very thing it was meant to measure.[9] Hájek makes essentially the same point:

> The problem is not with the agent assigning credences to her choices, but rather with the identification of credences with fair betting rates. This identification founders especially on cases in which the placement of the bet interferes with the proposition bet on. (p. 521, 2016)

Both Rabinowicz and Hájek focus on the identification premise, and reject it because measurements with bets can be disturbing. From a naturalistic perspective, in which we want to include the process of being offered the bet in the model (richness), they are correct to reject it.

On the surface, Joyce's response is a little different, as it focuses on deliberation. Joyce writes:

> [Levi] cannot prove anything about [the agent's] doxastic state *during* her deliberations by showing that she assigns extreme proba-

[9]Disturbing measurements of probability come up in other contexts, such as the standard theory of quantum mechanics (for example, see the description in Barrett (2019)). If one is not careful, the disturbing aspects of measurements can lead one to develop an entirely new theory, when the standard theory would suffice. Quantum probability theory is an example of such a new theory, which some have tried to apply to human cognition, going so far even to claim that the conjunction fallacy is not a fallacy because human cognition is non-classical (Pothos et al., 2017). However, once one incorporates into classical probability theory that the measurement is disturbing, there is no need of a new theory. Even those making the argument for quantum cognition note this: "Importantly, the idea that contextuality can change the meaning of superficially identical questions can be expressed classically too, since we can keep track of different meanings, through conditionalization" (p. 6, 2017).

> bilities to her actions after her deliberations have ceased. Friends of act probabilities can gladly grant that, once deliberation ends, [the agent] will be certain about both what she has decided and what act she will do as a result of her decision. But, since [the agent's] probability for A can (and usually will) *change* as a result of her deliberations, it is no news to be told that [the agent's] probabilities for A must be 1 *after* she decides on A. (2002, p. 83)

Joyce is pointing at the difference between an agent's degree of belief in an act A at the *end* of deliberation and *during* deliberation. On Joyce's view, the agent starts off uncertain about which act she will do, and through deliberation changes her probability.

The identification response and Joyce's response are deeply connected. Desirability tracking helps us see this. To see this, notice one further aspect of Rabinowicz's response: it fits perfectly into the desirability tracking account. His description of how the agent's probability for A changes upon learning that the bet is offered, if modeled fully (where the agent assigns prior probability to hypotheses about being offered gambles) is an example of desirability tracking.[10]

Desirability tracking precisely characterizes the way that probabilities change during deliberation, and highlights that the agent must be uncertain about *something* for her to still be deliberating. The reason that desirability tracking applies during the betting measurement procedure is that the agent is undergoing a kind of rational deliberation before she decides what to do. As

[10]Or, at the very least, could be made one. The agent would also need to be uncertain about whether or not she would learn that such a gamble was at play.

Rabinowicz describes it, the relevant piece of information that the agent hasn't yet learned is whether or not the bet will be offered. During deliberation—before the agent has learned all of the relevant pieces of information—she is uncertain about which act she will take. This shows us how Rabinowicz and Hájek's analysis is related to Joyce's.

The core point is this. Since being offered a bet is a disturbing measurement, when the bet is on something over which the agent has control, then we think of the agent as in the process of deliberation before she has learned all of the relevant information about the bet (including whether or not it will be offered). When the agent is offered the bet (or not!) this shifts her probability over what she will do: this is why it is a *disturbing* measurement. Desirability tracking characterizes how the probabilities evolve as the agent learns the relevant information. On this view, Levi correctly identifies that the agent *at the end of deliberation* knows what she will do but, like Joyce says, it does not show that this is so *during* the agent's deliberation.[11]

3.2.2 Betting rates cannot be applied

The second type of argument from betting dispositions is due to Spohn. He writes:

> The strangeness of probabilities for acts can also be brought out by a more concrete argument: It is generally acknowledged that subjective probabilities manifest themselves in the readiness to accept bets with appropriate betting odds and small stakes. Hence,

[11] I will discuss this further in §3.5.

> a probability for an act should manifest itself in the readiness to accept a bet on that act, if the betting odds are high enough. Of course, this is not the case. The agent's readiness to accept a bet on an act does not depend on the betting odds, but only on his gain. If the gain is high enough to put this act on the top of his preference order of acts, he will accept it, and if not, not. The stake of the agent is of no relevance whatsoever. (p. 115, 1977)

Rabinowicz makes Spohn's reasoning precise in the following way (2002). Consider a bet on some act proposition A, where two acts, A and B, are available to the agent. Let $U(A)$ and $U(B)$ be the (expected) value of A and B respectively. Let C represent the cost of the bet, and S the stake. Then, if the agent wins the bet, the agent receives $G := S - C$.

Thus, when the bet is in effect, if the agent takes act A, the expect utility is (now) $U(A) + G$ instead of just $U(A)$ (if the bet had not been placed). But then it follows that, if the agent is rational (and knows she is rational), she will take the bet on A if $U(A) + G > U(B)$. Similarly, she does not accept the bet if $U(A) + G < U(B)$.

This leads to a problem. Since the agent will accept any bet on A that has a $G > U(B) - U(A)$, we do not have any stable quotient C/S that determines the agent's betting behaviour. Rabinowicz writes, "Thus, the attractiveness of the bet does *not* depend on this quotient, but only on the value of G" (p. 99, 2002). But then, according to Spohn's premise that subjective probabilities manifest themselves in betting odds, since the quotient is equivalent to the betting odds, there can be no act probability.

There are two standard replies to this argument, made by both Rabinowicz and Hájek. The first is the same as the response to Levi's betting rates collapse argument: we need not identify probability with betting rates. Since this is a needed premise of the argument, rejecting it blocks the argument. As I described above, rejecting this premise is very sensible from a naturalistic point of view.

The second response is that Spohn's argument proves too much. Philosophers in the crowding-out camp tend to think that, even though deliberation crowds out prediction about one's *current* options, one can still reason about one's *future* options in the standard Bayesian way. However, Rabinowicz points out that (a strengthened version of Spohn's argument) rules out assigning probabilities to all of one's *future* acts as well:

> Let A be an action that will be available to the agent at some point in *the future*. Suppose that the expected utility of A is relatively low. However, if the bet on A is offered, the promised gain G, if sufficiently large, makes A an attractive prospect. But then, by the same reasoning as in the argument above, the agent's probability for A increases as soon as the bet offer is made. For the agent expects to perform A *if* he takes the bet. Therefore, and because the gain from the bet, if added to the original expected utility of A, would make A an attractive prospect, he expects to take the bet. Which means that the offer of a bet on A increases the probability of A for the agent. Consequently, for *all* future actions, their probabilities cannot be defined if such probabilities are supposed to correspond to the agent's betting rates. (p. 102, 2002)

Hájek says something similar about Spohn's argument:

> Nowhere does *deliberation* enter into this argument. So it should apply equally to your future acts that you are not deliberating about. (p. 522, 2016)

Hájek also thinks that the argument proves too much, as it also rules out future acts. However, he also takes future acts to be things about which one is not deliberating. A desirability tracking account of deliberation shows this to be the wrong analysis. The whole process of changing probabilities that Rabinowicz describes, which again is an instance of desirability tracking, *is* a kind of deliberation: the agent is learning things, and changing the probabilities across a partition in a way that tracks the desirabilities. From this view, the problem is not that there are some future acts about which we are not deliberating but for which Spohn's argument works: it is that the whole betting procedure on future acts *is* deliberation for the agent.[12] And as we have seen, desirability tracking is the fingerprint of deliberation.

[12]This framing also allows us to understand a point that Vavova (2016), following the analysis of Marušić (2015), makes against Hájek. Vavova argues that, even if deliberation doesn't crowd our prediction, it *complicates* it. In particular, in cases in which there is a tension between one's evidence, and one's view of oneself as a decision maker, then it is difficult to avoid irrationality. Vavova uses an example from Marušić (2015) of a marathon runner thinking about whether or not to run a marathon who has evidence that she might "wimp out" even should she decide to do it. This does indeed lead to complicated prediction—but this is just (plausibly) a case of weak desirability tracking! The complicated prediction just reveals that the marathon runner only has weak control over the partition. Just as with the case of betting, taking a more careful look at the example reveals that it fits a desirability tracking account. In this case, just not one of full control.

3.3 No Role for Act-Probabilities

Next I consider a further argument against act probabilities due to Spohn (1977).[13] The argument is that, since act probabilities play no role in decision making, we should eliminate them:

> First, probabilities for acts play no role in decision making. For, what only matters in a decision situation is how much the decision maker likes the various acts available to him, and relevant to this, in turn, is what he believes to be the result from the various acts and how much he likes these results. At no place does there enter any subjective probability for an act. The decision maker chooses the act he likes most—be its probability as it may. But if this is so, there is no sense in imputing probabilities for acts to the decision maker. For one could tell neither form his actual choices nor from his preferences what they are. Now, decision models are designed to capture just the decision maker's cognitive and motivational dispositions expressed by subjective probabilities and utilities which manifest themselves in and can be guessed from his choices and preferences. Probabilities for acts, if they exist at all, are not of this sort, as just seen, and should therefore not be contained in decision models. (p. 115, 1977)

As stated, it seems like Spohn is open to decision makers *having* probabilities over their own acts; just that they need not be included in the model.

[13] Levi also makes a similar argument (2007), and credits Spohn.

Rabinowicz points out that, in order to establish the stronger conclusion that Spohn seems to want in general,[14] he would need to show that act probabilities would be *harmful* to the agent (2002). Hájek notes the same thing, and points to Spohn's betting argument as trying to fill this gap (2016).

So, putting aside this objection that we could still add act probabilities if we wanted, let's focus on Spohn's claim that they play no role in decision making. Spohn is straightforwardly correct in Savage's decision theory: probabilities for acts do not enter into the value of an act all. However, this is not true in general. In Jeffrey's decision model—which I argued in chapter 1 dealt better with naturalism than Savage—probabilities for acts *do* factor into the expected value of an act, since we calculate the value of a proposition A as

$$U(A) = \frac{v(A)}{P(A)}$$

Having said that, it is true that, unless $P(A) = 1$ (which can't really happen in Jeffrey), the value of A will be the same.[15] So, while in a strict sense this is a role for act-probabilities, it is a fairly underwhelming one.[16]

Hájek points towards Skyrms' deliberational dynamics (1990) and Arntzenius' deliberational decision theory (2008) as examples of decision theories that make use of act-probabilities. While these do, I think that Spohn could point out that at least Skyrms takes his theory to be only qualitatively Bayesian.[17]

[14]Ruling out act probabilities.

[15]See for example p. 85 of *The Logic of Decision* (1983).

[16]Similarly, is also true that in some versions of causal decision theory there is a role for act probabilities. These are cases in which propositions about what you will do are correlated with dependency hypotheses (or partitions in a Skyrmsian K-parition). Hájek discusses this (2016.)

[17]I discuss this in more detail in §3.4. The relevant passage of Skyrms is on page 30 of

What Spohn would say about Arntzenius' theory, I am uncertain.[18]

Interesting, especially in light of the discussion to follow in § 3.4, Rabinowicz identifies the following possible role for act probabilities:

> In fact, they might do some good. According to Levi, I never deliberate whether to perform an option I am certain I am not going to choose. Now, something similar may apply to options with low probabilities: If I take an option to be very improbable, to begin with, then, in considering what to do, I might well allocate to it less time and effort in deliberation. (p. 113, 2002)

The idea is that, in addition to ruling out deliberation over probability 0 options, we might put less effort into learning about the desirability of low probability options.

Joyce's response to Spohn's objection is that while act probabilities *themselves* don't do anything useful, they arise from things that do. He writes:

> Why not abolish them? We now know the answer... We need act probabilities because (i) we need unconditional subjective probabilities for *decisions about acts* to *causally* explain action (though not to rationalize it), and (ii) we need Efficacy to explain what it is for an agent to regard acts as being under her control. (pp. 98-99, 2002)

The Dynamics of Rational Deliberation (1990).

[18]Rabinowicz (2002) makes a similar point about deliberation as a feedback process, referring to Gibbard and Harper's "Death in Damascus" example (1978).

Efficacy is a formal condition meant to capture the intuitive notion that an agent has control over a proposition. We can state Joyce's Efficacy condition as follows:

$$\text{Efficacy. } P(A \backslash dA) = 1$$

where A is some proposition, dA is the proposition "my deliberations will terminate in a desire to do A", P is the agent's probability function, and $P(\cdot \backslash \cdot)$ is the agent's *causal* conditional probability. Joyce uses the causal conditional probability because:

> The causal connection is the one that counts as far as questions of agency are concerned. Far from being a "metaphysician's plaything", causal probabilities are essential to understanding human agency. Unless we speak about [the decision maker]'s causal beliefs we cannot even say what it means for her to see herself as having a choice about A. (2002, p.89)

Efficacy, then, is supposed to determine the acts available to the agent. The propositions that satisfy Efficacy are those over which she has (or takes herself to have) a choice.[19]

It turns out that we "cannot have [the unconditional probabilities needed for Efficacy] without also having unconditional probabilities for A and $\neg A$. Act probabilities are not only coherent, they are *compulsory* if we are to adequately

[19]Really, if there is an act partition, we should require Efficacy to hold over each member of the act-partition.

explain rational agency" (p. 99, 2002).

Joyce's argument is the most similar to the response that a desirability tracking account of control lets us give to Spohn. According to the perspective developed in chapters 1 and 2, we do not *start off* with basic acts, which we can then exclude from the decision model. Rather, the agent's theory of the world, as given by her probability distribution over events, allows us to *identify* propositions that are sensitive to deliberation, and thus which we can consider to be acts (or, at the least, partially under the agent's control). Far from being insensitive to the distinction between things under the agent's control and things not, having things that the agent might think are under her control in the domain of her credence function is essential for a notion of control. From this perspective, we start off with probabilities over events, which lead to some being well-described as acts. Probabilities over acts come along for free, and do the real work of identifying acts.

3.4 The Argument from Vacuity

Levi is concerned that an agent who assigns probabilities to her own acts will have to assign them probability either 1 or 0, and that this would rob an agent of any decision at all (1991, 1993, 2007).[20] This last bit follows from a desire for the principles of rationality to be applicable non-vacuously. Levi expresses this in the following passage (which is one among many):

> When used for self policing, the applicability of the principles

[20]Both Skyrms (1984) and Jeffrey (1983) have expressed related worries. I will soon discuss these in more detail.

> should be nonvacous in the sense that a nontrivial distinction may be made between feasible options which are admissible for choice and others which are not. If the principles of rational choice never eliminate any feasible option from the relevant set of feasible options, they fail to serve this function. (pp. 25-26, 1997)

Here by *feasible* options Levi means an option that is possible for the agent to perform, and by *admissible* option one that is rationally permissible according to the agent's beliefs and desires. In a Bayesian framework, this would mean an option that maximizes expected utility.

This focus on nonvacuity is essential for Levi's argument against act probabilities to go through. Levi is careful to say that he is not arguing against act probabilities in general, but only under the assumption that the principles of rationality should be nonvacously applicable:

> I show that a rational decision maker assigning probabilities to hypotheses concerning the option he is about to choose must assign full belief to the prediction that the decision maker will choose rationally (i.e., choose an admissible option among those judged to be available) on pain of incoherence or contradiction. Only admissible options are seriously possible according to the decision maker. From this, according to the weak thesis, the admissible options and the feasible or available options must coincide. Hence, criteria for rational choice must be vacuously applicable in *all* contexts of choice. (p. 5, 2007)[21]

[21]Here the weak thesis is that:

Joyce helpfully breaks down Levi's argument into two stages:

> Premise-1: An agent who assigns probabilities to her present actions is required, on pain of irrationality, to assign a probability of zero to any inadmissible act.
>
> Premise-2: Once a deliberating agent assigns a subjective probability of zero to an action she no longer regards it as available for choice.
>
> ---
>
> Conclusion: An agent who assigns unconditional probabilities to her own acts cannot regard any inadmissible act as available for choice. (2002, p. 81)

Hájek breaks it down into even more fine-grained premises (2016), but here I will stick with the more coarse-grained version so that we can get a sense of the argument at bird's-eye view.

Premise-1 says that an agent who assigns probabilities to her own current acts *must* assign positive probability to only those acts that maximize expected utility. Levi in many places argues for this using the betting dispositions strategy discussed in §3.2 (for example, 1991, 2000, 2000). As in §3.2 I agree with Levi that *at the end of deliberation* an agent must assign probability 1 to the act she will do, and thus must assign positive probabilities to only admissible acts.

> In a situation of choice, an agent does not assign extreme probabilities, one or zero, to options among which his choice is being made. (p. 92, Rabinowicz (2002))

Levi is directly responding to Rabinowicz.

Even without the betting dispositions approach premise-1 is intuitive. For example, both Brian Skyrms and Richard Jeffrey have made the same claim. Similarly, both have thought that this can lead to issues for decision theory. Indeed, they also must endorse something like premise-2 as well, because they come to a similar conclusion as Levi. It will be helpful to consider Skyrms and Jeffrey's arguments.

To keep it distinct from Levi's charge against predicting one's own acts, I call the problem raised by Skyrms in *Pragmatics and Empiricism* the *problem of self-knowledge* (1984). Skyrms argues that, when an agent has too much self-knowledge, then she cannot be in a decision problem. Whereas Levi locates trouble with having credences over one's own acts, Skyrms locates the problem in the caliber of self-knowledge an agent has. Jeffrey raises a similar concern in *The Logic of Decision* (1983), and a related problem in Jeffrey (1977).

Skyrms raises the problem of self-knowledge in the context of a sophisticated decision maker in the spirit of Eells (1981). Skyrms phrases it nicely:

> Eells' decision maker is sophisticated indeed. He is not only rational but also knows that he is rational. He has degrees of belief about his degrees of belief and desires, and about acts and consequences conditional on his degrees of belief and desires. (1984, p. 73)

Eells introduced this level of sophistication in an attempt to reconcile evidential decision theory with causal decision theory. The details are unimportant here, but the strategy is not. The motivation for considering a sophisti-

cated agent is that when comparing decision theories, we should compare the strongest version of the two decision theories with each other. In particularly, we should not cripple the agents who are to use these theories. Thus, if it would be advantageous for an agent to condition on her beliefs and utilities, as Eells suggests, then we should allow her to do so. Note that this type of sophistication also fits with our idea of a naturalized agent. We want our agents to have degrees of belief about their acts, and other cognitive properties of themselves,[22] which is true of the sophisticated Eellsian agent.

Thus, for our purposes, we can consider Skyrms' argument to apply to a highly idealized and sophisticated naturalized agent so that we can understand decision theory in the most friendly of cases. Again, as stated in the introduction, no *real* agent can be fully Bayesian due to computational limitations, and no real agent ever considers a whole algebra as we imagine our agents do. When we relax these assumptions many more puzzles arise; here we are interested in puzzles that arise in the best-case scenario of an idealized agent who considers propositions about her own acts, beliefs, utilities etc.

Skyrms makes the following remark (1984):

> The theory must walk a tightrope between too much self-knowledge and too little... If the individual has too much self-knowledge, then he already knows what he will do and there is no decision problem (i.e., his prior probability that he will choose a certain act is one.) (p. 74)

[22]Since our model of their decision making satisfies richness, agents will entertain hypotheses about themselves.

This is similar to the following remark of Jeffrey about how an agent reasons about her acts in a decision problem (1983):

> Then the act $\bar{W}$ of bringing red wine is preferred. If W were totally within the agent's power to make happen or not, we should then have $prob(\bar{W}) = 1$, since the preferred act will be performed. (p. 85)

Similarly,

> Suppose you think it within your power to make A true if you wish, that you prefer A to W, and that you are convinced that A is preferable to every other one of your options. Then $P(A) = 1$, for you know you will make A true. (p. 136, 1977)

Both Skyrms and Jeffrey claim that an agent (satisfying some background conditions of self-knowledge) must already know what she will do.[23]

[23] Another person who makes this claim is Sneed (1966). He writes:

> Suppose that the agent believes that he can make any member of [the decision partition] he chooses true. Then it seems natural to assume that, after he has completed his deliberation, the proposition picked by the deliberation will have a probability near one, while the others will have a probability near zero. The use of "near one" and "near zero" is justified by appealing to the claim that the agent really only believes he can try, with very high, yet not certain, chances of success, to make the member [of the decision partition] he chooses true. (p. 272, 1966)

In the last bit about "trying" Sneed is following Jeffrey's suggestion in *The Logic of Decision* (1965). Of course, according to a desirability tracking account, the probability of a proposition need not be 1 for the agent to no longer view herself as in a decision problem. All that is needed is for the partition to, at that point in time, cease to be desirability tracking. Thus the probability 1 aspect of all these claims makes understanding the point easier, but is not necessary.

Importantly, even though Levi uses the betting dispositions strategy to establish premise-1, this exact approach is not required. Skyrms' and Jeffrey's reasoning does not appeal to betting dispositions. Nor does it need to. Frederic Schick has made this reasoning very precise (1979):

Schick writes

> I have been speaking of this as a problem of the foreknowledge of choices, but knowledge comes into it only indirectly. Basically, it is a problem of *forebelief*. Let me put it more fully. Suppose that the agent thinks that his [decision problem] will be made up of options $o_1, o_2, \ldots o_m$, and that [he has degrees of belief about the consequences of the different options]. (An *option* here is a *live* option: not just anything the agent might do, but one of a set of possible actions that, conjointly, raise [a decision problem] for him.) Let the agent also think that his overall preference ranking will be R.... Let the agent believe that he will choose rationally—that he will choose an option whose expected utility won't be exceeded by that of any other—and let him believe that the choice he will make will be very effective. Suppose that he will not change his mind on any of this before he chooses. It follows from what he believes that some set S of his options is the set from which he will choose one, and that this option will come out. To simplify a bit, suppose that S contains only the singleton option o^x. The agent is now committed to believing that he will do x [where x is the actual event that his decision leads to]. He will also be bound to believe this at the point of choice. If he does believe it then, he

> won't have any issues of whether to do x or not, and so won't have any choice to make. Foreseeing his choice-situation precludes his having a choice. (p. 239, Schick (1979))

He has also generalized it so that it does not depend on the agent believing that she will choose *rationally*, but only that she will choose in some particular way. Even if the agent just has a theory of her own choice rule, as Schick writes, the "problem remains: foresight undoes choosing" (p. 240, 1979).[24]

We also see that Schick's argument doesn't just get us premise-1, but also premise-2 and all the way to Levi's conclusion. This leads Schick, like Levi, to reject act probabilities:

> We have seen that logic alone rules out our knowing the whole truth about ourselves. Where we will (effectively) choose during t, self-omniscience is out during t. This is the basic conclusion above. (p. 243, 1979)

This worry about self-knowledge is also very Skyrmsian. Skyrms advocates moving from a static version of decision making to a *diachronic* version of decision making (p. 74, 1984). The particular version of dynamic deliberation Skyrms considers in *Pragmatics and Empiricism* (1984) is what he explored more fully in *The Dynamics of Rational Deliberation* (1990). There, he is very clear that in his account of deliberation he is not considering a full-blooded Bayesian agent:

[24]For the full description of Schick's reasoning, see pp. 241-243 of *Self-knowledge, Uncertainty, and Choice* (1979).

> We assume that [the agent] moves according to some simple dynamical rule for "making up one's mind," as opposed to performing an elaborate calculation at each step. This rule should, however, be "qualitatively Bayesian" in various ways. It should reflect her knowledge that she is an expected utility maximizer and the status of her present expected utility values as her expectation of her final utility values. (1990, p. 30)

So Skyrms' deliberating agents are *qualitatively* Bayesian, and not ideally Bayesian like Levi and Schick's agents. Skyrms' agents avoid Levi and Schick's concerns about foresight ruling out choice.

Here, we want to know if fully-blooded Bayesian agents can also avoid the problem. I'll ultimately argue that they can, also by taking a diachronic perspective, as informed by the desirability tracking account of control I provided in the previous chapter. Before I sketch how this works, let us first consider the arguments the pro-act probability folks make against Levi.

Here I'll focus on the arguments that Joyce (2002) and Hájek (2016) make against Levi. I will show how their responses violate the austerity condition of naturalism: they try to avoid Levi's conclusion by adding additional structure into the agent's model, which goes beyond the agent's theory of the world.

First I'll focus on Joyce's response. I've already discussed his response to Levi's premise-1.[25] Here I'll focus on premise-2. Recall that premise-2 was

> Once a deliberating agent assigns a subjective probability of zero

[25] In §3.2.

to an action she no longer regards it as available for choice.

Joyce takes the following passage of Levi's as evidence that this is Levi's view:

> if [the agent] is convinced that [she] will choose optimally, [she] is convinced that every proposition describing a suboptimal course of action will be false. Suboptimal courses of action will have been ruled out as possibilities and, hence, as available options for choice. (2000, p. 394)

Joyce makes two important clarifications about this passage. First, he notes that Levi uses "possibility" in an epistemic sense, not any kind of metaphysical or logical sense (2002). Second, he points out that Levi must be referring to acts in a *de re* sense, not a *de dicto* sense. For, on the latter reading, the statement is trivially true: the agent "would merely be convinced that *whatever* act she ends up choosing will be rational" (2002, p. 87).

Having made this clear, Joyce identifies the different answers to the following question as the crux between him and Levi: "does the fact that [the agent] is certain that she will not perform a given act prohibit her from seeing that act as available for choice?" (2002, p. 87) Levi thinks yes,[26] whereas Joyce thinks not. We can understand this disagreement best by understanding it as a disagreement about the relationship between *epistemic possibility* and *practical possibility*, where the meaning of these terms is given by the following definitions:

[26] And Schick, and Skyrms, and Jeffrey.

> H is a serious *epistemic* possibility for [an agent] iff it is consistent with the laws of probability that she assign H a positive probability while assigning probability 1 to each proposition in her corpus of certainties.
>
> ...
>
> H is a serious *practical* possibility for [an agent] iff she is rationally required to factor the possibility of H's truth into her decision making. (2002, p. 90)

Levi's view is that, if an act is *not* a serious epistemic possibility, then that act is not a serious practical possibility. Joyce's view is that there are some cases in which there are acts that are not serious epistemic possibilities but *are* serious practical possibilities.

Skyrms' problem of self-knowledge relies on reasoning very similar to Levi's: an inference from epistemic impossibility (captured by the idea of assigning probability 0 to all acts except one) to practical impossibility (not needing to consider these acts because they have probability 0). This is what leads to a lack of decision problem—since every other act gets probability 0, there is nothing more to deliberate about.[27]

Joyce's argument against this inference from serious epistemic impossibility to serious practical possibility rests on a distinction between conditions that are external to the agent and those that are internal. He writes,

> If A or dA[28] is epistemically impossible because [the agent] is cer-

[27] This is also what Schick relies on—see how he derives (11) on p. 241 (1979).

[28] Where dA is the statement "My deliberations will terminate in a decision to A" (2002,

> tain of some *exogenous, uncontrollable* condition H that is incompatible with A or dA, then A really is a dead issue for her... Levi thinks the same holds when a person becomes certain of facts *internal* to her deliberations.

Joyce's response depends on this internal/external distinction. But by the naturalism developed in §1.2, it is unclear what would support such a distinction. The agent reasons about facts about herself, just like she reasons about anything else. If she believes with probability 1 a proposition is false, then she does not need to consider it in her deliberation. The inference from epistemic impossibility to practical impossibility is valid.

To make this clear, recall the Skyrms' quote from §2.3.1:

> Indeed, one can argue that if a deliberator is absolutely sure which act he is going to do he needn't deliberate, and if he is absolutely sure he won't do one of a set of alternative acts his deliberations should concern only the others. Putting it the other way around, if a decision maker thinks that there is any chance that deliberation might change his probabilities of an act, he should have given the act a probability different from zero or one. (p. 36, 1990)

This last bit is essential. Skyrms is saying that, if a decision maker believes that she might change her mind about an act, then she should not be certain that she will perform (or not perform) that act. This follows from standard

p. 87).

Bayesian reasoning: if the decision maker thinks it is possible that her deliberation will change the desirability of an act (in a decision-relevant way) before the time of decision, then she will *not* have assigned that act probability 1 or 0.

Joyce's argument relies on the agent having some primitive ability to change one's evidence, when it comes to one's actions. He writes:

> [The decision maker] *controls what evidence he has concerning his own actions.* No external obstacle prevents him from changing his mind, and by doing so he can alter the constitution of his corpus of certainties. If he changes his mind and decides on A, then dA will replace $d\neg A$ in his corpus, and this both destroys his evidence for $\neg A$ and gives him evidence for A. Indeed, this evidence for A is conclusive so long as he takes his decisions to be causally efficacious. The point is that, *insofar as A and $\neg A$ are concerned, the [decision maker] controls the contents of his corpus of certainties and so controls what it is reasonable for him to believe about A and $\neg A$.* Given this, the mere fact that A conflicts with his evidence cannot rule it out as a serious practical possibility.... This means that A is practically possible even though [the decision maker] is sure he will not change his mind and perform it. (p. 92, 2002)

But this simply posits some basic ability to the agent, even if there is a complicated story about evidence in the background. It says that, *even if she is certain* that the world is not a certain way, she still *could* make it that way. Why would such a thing be true? Joyce is relying on an idea of controlling

one's evidence regarding one's acts. It certainly violates the austerity condition of naturalism: it relies on additional structure in the agent's attitudes, that she expects to have no empirical consequences (since she assigns such a belief change probability zero).

Furthermore, as I showed in chapter 2, we have available to us a naturalized version of control (desirability tracking) that avoids any reference to basic powers. So, not only does Joyce's control violate naturalism by positing something dangerously close to libertarian free will, we have a notion of control available to us that avoids this defect.

Recall one of the motivations for endogenizing control: we wanted a notion of control to come from the agent's theory of the world—how she fits into things. Desirability tracking does exactly this. Informally, it says that an agent views a proposition as under her control if, as she learns things that make that proposition more desirable, the probability of that proposition moves in the right direction. Positing basic powers, where the agent needs to do a non-Bayesian update *that she believes she will never do*,[29] does not arise from the agent's theory of the world.

Moving to Hájek, he is even more explicit:

> Even when you are certain that you *will not* perform a given action, you may well be *able* to perform it. Even when you are certain that you *will* choose the red wine, you are still *able* to choose the white.

[29]Joyce writes: "On the other hand, [the decision maker] might simply delete $d\neg A$ from his corpus of certainties" (p. 93, 2002). This is the non-Bayesian update. Furthermore, the agent must assign probability 0 to doing such an update, otherwise, by reflection, $d\neg A$ would not be in his corpus of certainties. Compare this to the Skyrms quote above.

(p. 520, 2016)

This is basically just postulating some kind of basic powers. Again, we'd want a theory of control or action to come from the agent's theory about how the world works. Desirability tracking gives us that. Hájek's account just puts it in.[30]

This might be driven by intuitions about freedom. Even though I know (believe with probability 1) that I *won't* do something, doesn't mean that I *can't*. This would mean that, pace Levi, even *at the moment of choice*, after she has factored in all of her information and believes with probability 1 that she will do something, the agent could do otherwise. Joyce and Hájek are trying to save this intuition. But Hájek at least wasn't comfortable with this strategy (arguing from freedom) in other contexts. Hájek writes that those who argue against act probabilities sometimes appeal to "Wishy-washy agency/free will argument[s]" (p. 514, 2016). But this seems to be exactly what just positing basic powers is.

Joyce's response is more subtle and clear: he is very explicit about how we cache out these powers formally. Joyce, following Velleman (1989), appeals to

[30]Hájek also has other arguments against Levi's argument from vacuity, but they seem to miss the mark (2016). His first argument (p. 519) argues that, when acts are tied for choice-worthiness, then the principle of choice is applied nonvacously. On a desirability tracking account it is not possible to have options that are tied, since deliberation won't affect their probabilities. Also, even if his argument worked in the case of ties, it would be a pretty sorry conclusion if choice was only possible when it didn't matter. His second argument is that rational agents need not be *smug* in the sense of being certain they are rational. Again, it would be a sorry conclusion if choice were impossible for agents who know they are rational. Furthermore, Schick (1979) shows that agents don't need to believe they are rational for this argument to go through, but only that they need to know their choice rule. His remaining argument is that during deliberation the agent hasn't yet ruled out the admissible options. This is basically the position I am endorsing, except I don't view it as showing that Levi is wrong *about the moment of choice*.

the idea of self-fulfilling beliefs, as cached out in terms of something he calls *Efficacy*, which is a causal notion.[31] Efficacy ensures that the agent's decision to do A (dA) ensures that A in fact obtains. Joyce holds that an agent can believe this, even when she assigns probability 0 to dA obtaining.

Joyce's view has the virtue that it is very explicit about the extra structure that must be added to the agent's conception of the world: basic causal properties that obtain even when an agent is certain the cause (dA) will not obtain.[32]

Before I close out this section I want to return to the definition of a serious practical possibility. Recall that something is a serious practical possibility for

[31] Joyce does consider what he calls a "deeper worry" (p. 94, 2002), which is that he might be accused of "portraying the agent who changes her mind as altering her *beliefs* about what she will decide *on the basis of no evidence whatever*" (p. 94). His response is that, because of the self-fulling nature of such beliefs (for example, about some proposition H), they "involve being in a position to disregard evidence concerning H, because one knows that it will be made moot by the fact of one's belief" (p. 96). From the Bayesian perspective on evidence, it is unclear what to make of this statement. Perhaps it means that things people might *typically* take to be evidence are *not* evidence in this case, given the agent's beliefs. That is fine. But one doesn't get to ignore *evidence*, if evidence means evidence from the agent's own point of view. Ultimately, the agent thinks that either H is true or H is false. She might also believe that her belief in H somehow effects H in a way that is self-fulfilling. In the case of self-fulfilling belief, I agree with Joyce when he says that the agent's confidence in dA will "wax or wane *in response to information about A's desirability relative to her other options*" (p. 97). But I am not sure why this leads to the agent being able to ignore *any* evidence—desirability is evidence in this case! And this is exactly what a desirability tracking account of control relies on. But it does so without these special powers to remove things from one's corpus of certainties. So desirability tracking agrees with much of Joyce's reasoning, without endorsing the conclusion that the agent views herself as being able to perform probability 0 events.

[32] This is similar to another way one might want to ground such basic powers: through counterfactuals. For example, Jennie Louise writes:

> Here, it is important to note that when we say that an agent *won't* Φ, we are not asserting that they *can't*: in other words, the assertion that the agent lacks the capacity for Φ-ing is not usually the basis for the assertion that they won't Φ... And we can cash this out as follows: where an agent won't Φ even though they can, this means that, while there is little probability that the agent will Φ in the actual world, there is a range of relevantly similar possible worlds in which the agent systematically *does* Φ. (pp. 333-334, 2009)

Even though Louise allows positive probability for the agent to Φ in the actual world, one could suppose it is 0 and still try to cash out possibility in terms of counterfactuals.

an agent if she must factor the possibility of its truth into her decision making. Notice that in expected utility theory proper, probability 0 events cannot be serious practical possibilities: they contribute nothing to the expected value of an act. Furthermore, in Jeffrey's theory, probability 0 events aren't even assigned desirabilities. Thus, on the face of it, it seems that probability 0 events will not be serious practical possibilities for agents in general, and especially for Jeffrey-style agents.

Here is where we stand: Levi's argument seems to go through, unless one adds additional structure like Joyce and Hájek want to do. This structure all lives in probability 0 events, and, in this sense, violates the austerity condition of naturalism. It is also unclear how it fits with Jeffrey's framework, since it uses a strictly positive probability measure.

3.5 The Ridge

Now we have surveyed both camps. I have shown how neither camp satisfies naturalism. The camp that wants to outlaw act probabilities violates *richness*. They leave things (propositions dependent on an agent's acts) out of the domain of an agent's belief. The camp that welcome act probabilities violates *austerity*. They add additional structure beyond the agent's theory of how the world *actually is* by appealing to some kind of basic powers (Hájek) or causal/counterfactual structure (Joyce, Louise).

My ridge does neither. On my ridge, we have an account of how probabilities over an event space give rise to a certain interplay between an agent's

degrees of belief and an agent's desirabilities. This interplay, characterized by different types of desirability tracking, gives an agent a way to view herself as deliberating.

Let us recall the goal of this thesis: we wanted to naturalize decision theory. That is, we wanted to build a model of an agent in which she can view herself as both agential and part of nature like everything else. When an agent's credences across a partition satisfy some version of desirability tracking, she views her own learning as leading to better outcomes. In this way, we have resolved the puzzle. On this ridge we are, like Caf's daughter Asia in the epigraph, less filled with woe, and more with thought, imaging how our cognition in the world might lead to future glory.

Why, then, has Eris had so much success at dividing decision theorists among the two camps? And why has her work pushed them to various stripes of non-naturalism?

The core, from my vantage point on the ridge, comes down to the various decision theorists feeling a threat to *agency*. Connected to agency is a sense of *freedom of choice*. Even if I *in fact* do A, I *could* have done $\neg A$! This is the intuition people are trying to save. Spohn, in a very clear and honest article, writes about the desire to view oneself as having free will (2007): "Everyone sober believes in freedom of the will" (p. 297). In this context he points out a threat:

> The world, which includes us, may be deterministic. This generates a contradiction. Or the world may be indeterministic, but it seems common ground that this does not improve the dialectical

> situation; the freely willed cannot occur at random. There is no escape; we are dealing here with a sharp contradiction, an outright antimony. (p. 297, 2007)

Spohn's strategy is to "plead for compatibilism, but only by distinguishing two perspectives" (p. 300). He does this *by leaving the agent out of the model*—by violating richness. This allows one to view one's acts as free:

> ...this *no probabilities for acts* principle entails the *acts are exogenous* principle, which says that the possible actions of issue in a decision model are exogenous, that is, *first causes* or *uncaused* within the model.

This is a very honest assessment of why one might want to leave out act probabilities from a decision model: they threaten free will. Spohn is very clear that from an *empirical* perspective on oneself, of course one will assign probabilities to one's own acts. But then this will not fit your desire for free will: "If you search for it only within the empirical perspective, you are lost in paradox" (p. 300).

This move from a third person (empirical, naturalistic) perspective to a first person (normative, free-will) perspective is shared by Levi:

> The thesis I wish to advance is that this demand for nonvacous self applicability entails an asymmetry between the first person perspective and the third person perspective which has no bearing on first person privileged access but which does pose a serious

> obstacle to viewing principles of rational choice designed to be nonvacously applicable in self criticism as generalizations useful in prediction and explanation of human behavior. (p. 26, 1997)

If being an agent means being able to "deploy criteria for choice to determine what he should do" (p. 32, 1997), then we are led to Levi's worry that to "be an agent crowds out being a predictor" (p. 32). That is, like Spohn, we can preserve some notion of agency if we do not consider ourselves in the model: "[we] must not make any judgements as to the probability as to what [we] will do" (p. 32).[33]

So, one broad strategy to save agency is to leave the agent out of the model, thus violating richness. Spohn points us towards a second strategy:

> The large majority seeks to realize compatibilism by saying that an action is free if and only if it is *appropriately* caused, and then all effort goes into explicating what "appropriately" is to mean here. (p. 303)

This is essentially the strategy I pursue with desirability tracking: it specifies the correct relationship[34] between beliefs and desires, that allows one to view oneself as agential. I prefer not to use the language of *free* action, and instead of

[33]Both Spohn and Levi are to some extent driven by the Kantian view of autonomy. For example, Levi writes: "Still, if there is any remnant of the Kantian view of autonomy worth preserving from a pragmatist perspective, it is to be found in the nonvacous applicability of standards of rationality in self criticism" (p. 38, 1997)

[34]As stated, desirability tracking is noncausal. However, one could modify it to make it causal by changing using the conditional *causal* probability in many of the statements, as Joyce does in Efficacy. This would lead to a version of desirability that says something like, an agent views herself as having control over something if her learning something *causes* the world to evolve in a certain desirable direction.

control, because (as I believe this chapter illustrates), intuitions about freedom can lead us astray.[35]

Spohn rejects any strategy along these lines because he has commitments concerning other spheres of human life (moral and legal), and the account of freedom recovered by the strategy above does not satisfy his commitments.[36] However, for those like me who do not share such commitments, the desirability tracking strategy I provide recovers an adequate notion of agency, or freedom, if one wishes to call it that.

Why, then, all the extra structure from the pro-act probability camp? If we have, with desirability tracking, an account of control that satisfies naturalism, and which allows an agent to view herself as agential, why do we need the additional structure?

I think here we hit issues about the difference between *deliberation* and *choice*.

[35]We might understand this as a decision-theoretic sharpening of some classical compatibilist ideas. For example, Paul Russell writes (2021),

> According to the classical compatibilist strategy, not only is freedom compatible with causal determinism, the absence of causation and necessity would make free and responsible action impossible. A free action is an action caused by the agent, whereas an unfree action is caused by some other, external cause. Whether an action is free or not depends on the type of cause, not on the absence of causation and necessity.

This kind of view is often attributed to Hume (due largely to the section "Of liberty and necessity" in *A Treatise of Human Nature* (1896), who is considered one of the most influential of the compatibilists.

[36]Spohn writes:

> Even more directly, the (first-order desires) must have the right kind of content. They must conform to moral duty (and the categorical imperative), or they must be humanely adequate in respecting our rational nature or in perfecting our virtues. These and other ideas are discussed with great care in the philosophical literature. In all this, one must never forget that an appropriate notion of freedom goes hand in hand with appropriate notions of human dignity, responsibility, and blameworthiness with all its practical, moral, and legal implications. (p. 303, 2007)

Desirability tracking gives one an account of rational deliberation. There is not, however, a moment of choice where *in fact* one might have done otherwise.

This is one point on which I think Levi is absolutely clear, and yet has not been adequately addressed in the whole debate. This is the idea that act probabilities are only permissible when the acts are *currently* available to the agent—what Levi calls the "Moment of Truth" (p. 33, 1997). He spells this out more fully (when describing an example decision maker, Sam):

> These remarks reply, as I have said, to Sam's judgements (or, for that matter, to the judgements of any deliberating agent X) at that stage in deliberation where the agent has identified his values, convictions, and options sufficiently to apply the principles of rationality to the evaluation of the admissibility of these options. They do not apply to Sam's evaluation of the rationality of Y's choices or to the rationality of Sam's choices in some other future context of deliberation. (p. 33, 1997)

Levi, in all his discussion, is clearly considering the moment when the agent *has* all of the evidence integrated into her beliefs—right at the *moment* of decision.[37] Again, this is the same kind of point that both Skyrms and Jeffrey make.

Jeffrey's solution, and mine, is to go explicitly dynamic:

> And even if you are convinced that (say) your options are A, W,

[37]Much of the discussion ignores this. For example, recall Hájek's statement about him being confident, but not certain, that he will have some wine. This is a statement about a *future* decision.

> and $\neg A$, and that A is fully in your power to make true, it may be that $P(A) \neq 1$ because you consider that before the time comes to enact A, your preferences may change for some reason or other so that when the time comes, A may no longer be the highest-ranked of your three options. (p. 137, 1977)

But neither Jeffrey nor I really disagree with Levi's main point: *at the Moment of Truth*, "if your preferences *at that time* are given by P, U, *then* $P(A) = 1$" (p. 137, Jeffrey, 1977).

If one agrees with Skyrms that when an agent already knows what he will do, then this means that "there is no decision problem (i.e., his prior probability that he will choose a certain act is one.)" ... (p. 74), then we have Levi's point: there is never *really* a decision problem—never a moment of *genuine* choice, where the admissible actions are a proper subset of the feasible ones.

From my naturalistic perspective, there is no problem. I have a theory of control, and deliberation, and this is enough to get me a (perhaps weak) kind of agency. Indeed, I think it is the only kind of agency compatible with he kind of naturalism I sketched in chapter 1.

However, if someone *really* wanted an account of agency that *had* a moment of genuine choice, and this person wanted to include propositions about the agent in the model, then they might do so by adding additional structure. This is exactly what Joyce and Hájek do. Unsatisfied with a view of agency without a moment of choice, they opt to include some kind of additional structure in the model that allows for the agent to do things that she is *certain* she will not. In this way they violate austerity: they add things that go beyond the

agent's theory of how she thinks the world actually is.

From my version of naturalism, it is unclear why we want this. Desirability tracking allows us to identify meaningful deliberation. In this sense, we have a naturalistic notion of agency. If I don't ever view myself as making a *genuine* choice, where I really could have done otherwise, then that is no problem.[38] It won't change how I expect the world to go. In fact, I am *certain* that it won't.

Perhaps, following Levi, the thing one loses is a view of oneself as what he calls a rational agent, in favour of a view of oneself as what he calls a rational automaton:

[38]Of course, sometimes we *do* speak as if we really could have done otherwise. About those situations, I would say something like this. Suppose, for example, I notice my friend make what I take to be a wrong move in a board game, but I believe my friend is rational, I think she had different priors than I did over relevant propositions. Suppose I think I am better informed. Insofar as I think my friend might encounter similar situations again, I might say to her something like, "I think you made the wrong move. You played A, but you should have played B, because of fact C". Really, what I am saying is that my friend's P and U as they were deliberating might not have conditioned on C, which would have led her desirability to flip A and B in the ranking. Then, since my friend now is aware of C, next time she is in a similar situation she can do B. So this is all forward looking: I am pointing out some fact (C) that I think will flip my friend's ranking of future acts in similar scenarios, and I can do this without really believing that she could have done otherwise. She might then defend A by saying she was aware of C, but that also D matters and in this case that flipped them back, and so on.

Another example: you face a sequence of decision problems, in which you can choose to enter room A or room B (then A' and B', and so on). You don't get to see inside before you enter, but you know that one room has &0 and the other has $100, you just don't know which. After you enter the room you learn what is in the room (and thus the other as well). Then you move to the next choice, between rooms A' and B', and so on. We might say something like, "I made the wrong choice first, because A had $0 and B had $100. If I had gone in room B, I would have been richer!". But, given your priors, you made the right choice. I think why we tend to do this is because we are really using the learning about the outcome to inform our future choices. Say, for example, that the distribution of money between the rooms on one of the decision problems is dependent on the previous ones (you view what was in A as indicative about what will be in A', for example). Then I am learning about the expected value of the rooms for future decisions. I view A and A' as relevantly similar, and the counterfactual talk about that decision is really an artifact about how I am thinking about future decisions. The core idea is pragmatic: the counterfactual talk is a heuristic that approximates the learning I should actually be doing about future events.

> External policing can take the form of deploying norms of rationality as blueprints for rational automata. There is no clash between using principles of rationality for explanatory and predictive purposes, on the one hand, and using them prescriptively for designing rationally acceptable conduct. Tension arises only if, in addition to using them for external policing, one seeks to use them for internal policing. In that event, the blueprints can no longer be for rational *automata*. Agents will satisfy the requirements for rational health only if they apply the principles of choice to evaluate their options. But, in that case, neither they nor we, the outside agents, can regard them as predicting their own choices. Rational automata can predict their own choices. Rational agents cannot. (p. 37, 1997)

Perhaps Hájek and Joyce, and some others in their camp, following Levi, want to view themselves as rational *agents* as opposed to rational automata. However, instead of violating richness, they also want to predict their own acts. So they violate austerity, by supposing that they can do something that is epistemically impossible for them.

I have illuminated this middle ridge. I have argued that it takes seriously the strongest arguments against act probabilities. However, I have shown how it also makes room for deliberation, and for assigning probabilities to every event in an agent's algebra. Perhaps, on my ridge, we only view ourselves as rational automata, and not rational agents in some very strong sense of agent. Perhaps that is the price of naturalism. Maybe, now with this ridge visible from both camps, some philosophers will join me here. We have act probabilities, we have agency in the form of desirability tracking, and we have naturalism. Perhaps

soon we will have the pleasure of your company as well.

***

Epilogue

> Anon rush'd by the bright Hyperion;
> His flaming robes stream'd out beyond his heels,
> And gave a roar, as if of earthly fire,
> That scared away the meek ethereal hours
> And made their dove wings tremble. On he flared...
> — John Keats, *The Fall of Hyperion*

Bibliography

Ahmed, A. (2014). *Evidence, decision and causality*. Cambridge University Press.

Aristotle (1926). *Nichomachean Ethics*. Loeb Classical Library. Harvard University Press.

Armendt, B. (1993). Dutch books, additivity, and utility theory. *Philosophical Topics 21*(1), 1–20.

Arntzenius, F. (2008). No regrets, or: Edith piaf revamps decision theory. *Erkenntnis 68*, 277–297.

Barrett, J. A. (2019). *The conceptual foundations of quantum mechanics*. Oxford University Press.

Bolker, E. D. (1967). A simultaneous axiomatization of utility and subjective probability. *Philosophy of Science 34*(4), 333–340.

Bolker, E. D. (1974). Remarks on'subjective expected utility for conditional primitives'. *Balch, McFadden and Wu (1974) 79*, 82.

Broome, J. (2017). *Weighing goods: Equality, uncertainty and time*. John Wiley & Sons.

Christensen, D. (1996). Dutch-book arguments depragmatized: Epistemic consistency for partial believers. *The Journal of Philosophy 93*(9), 450–479.

Colvin, S. (1925). *John Keats; His Life and Poetry: His Friends, Critics, and Afterfame*. Scribner's.

Demski, A. and S. Garrabrant (2019). Embedded agency. *arXiv preprint arXiv:1902.09469*.

Diaconis, P. and B. Skyrms (2017). *Ten great ideas about chance*. Princeton University Press.

Eells, E. (1981). Causality, utility, and decision. *Synthese*, 295–329.

Eells, E. (1982). *Rational decision and causality*. Cambridge University Press.

Finetti, B. d. (1992). Foresight: Its logical laws, its subjective sources. In *Breakthroughs in statistics*, pp. 134–174. Springer.

Fishburn, P. C. (1964). *Decision and value theory*. New York: Wiley.

Gibbard, A. and W. L. Harper (1978). Counterfactuals and two kinds of expected utility. In *Ifs: Conditionals, belief, decision, chance and time*, pp. 153–190. Springer.

Hacking, I. (1967). Slightly more realistic personal probability. *Philosophy of Science 34*(4), 311–325.

Hájek, A. (2016). Deliberation welcomes prediction. *Episteme 13*(4), 507–528.

Hild, M. (1998). Auto-epistemology and updating. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 92*(3), 321–361.

Hintikka, J. (1979). Impossible possible worlds vindicated. In *Game-theoretical semantics*, pp. 367–379. Springer.

Howson, C. and P. Urbach (1993). *Scientific reasoning: the Bayesian approach*. Open Court Publishing.

Hume, D. (1896). *A treatise of human nature*. Clarendon Press.

Huttegger, S. M. (2013). In defense of reflection. *Philosophy of Science 80*(3), 413–433.

Huttegger, S. M. (2014). Learning experiences and the value of knowledge. *Philosophical Studies 171*, 279–288.

Huttegger, S. M. (2015). Merging of opinions and probability kinematics. *The Review of Symbolic Logic 8*(4), 611–648.

Hutter, M. (2004). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media.

Ismael, J. (2007). *The situated self*. Oxford University Press.

Ismael, J. (2016). *How physics makes us free*. Oxford University Press.

Jeffrey, R. C. (1965). *The logic of decision*. University of Chicago Press.

Jeffrey, R. C. (1977). A note on the kinematics of preference. *Erkenntnis 11*(1), 135–141.

Jeffrey, R. C. (1983). *The logic of decision*. University of Chicago press.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Joyce, J. M. (2002). Levi on causal decision theory and the possibility of predicting one's own actions. *Philosophical Studies 110*(1), 69–102.

Joyce, J. M. (2020). Yet another refutation of causal decision theory?

Krakauer, D., N. Bertschinger, E. Olbrich, J. C. Flack, and N. Ay (2020). The information theory of individuality. *Theory in Biosciences 139*(2), 209–223.

Krantz, D. H. and R. D. Luce (1974). The interpretation of conditional expected-utility theories. *Essays on Economic Behavior under Uncertainty*, 70–73.

Kyburg, H. E. (1978). Propensities and probabilities. *Dispositions*, 277–301.

Levi, I. (1967). Probability kinematics. *The British Journal for the Philosophy of Science 18*(3), 197–209.

Levi, I. (1991). Consequentialism and sequential choice. In *M. Bacharach & S. Hurley (Eds.) Foundations of decision theory: Issues and Advances*, pp. 92–122. Oxford: Basil Blackwell.

Levi, I. (1993). Rationality, prediction, and autonomous choice. *Canadian Journal of Philosophy Supplementary Volume 19*, 339–363.

Levi, I. (2000). Review essay: The foundations of causal decision theory. *Journal of Philosophy 97*(7), 387–402.

Levi, I. (2007). Deliberation does crowd out prediction. *Homage a Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz. E*.

Levi, I. et al. (1997). *The covenant of reason: rationality and the commitments of thought*. Cambridge University Press.

Lipman, B. L. (1991). How to decide how to decide how to...: Modeling limited rationality. *Econometrica: Journal of the Econometric Society*, 1105–1125.

Louise, J. (2009). I won't do it! self-prediction, moral obligation and moral deliberation. *Philosophical Studies 146*, 327–348.

Marušić, B. (2015). *Evidence and agency: Norms of belief for promising and resolving*. Oxford University Press, USA.

Miller, B. E. (1965). On the incompleteness of keats' "hyperion". *CLA Journal 8*(3), 234–239.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In *Essays in honor of Carl G. Hempel*, pp. 114–146. Springer.

Pettigrew, R. (2021). Logical ignorance and logical learning. *Synthese 198*(10), 9991–10020.

Pothos, E. M., J. R. Busemeyer, R. M. Shiffrin, and J. M. Yearsley (2017). The rational status of quantum cognition. *Journal of Experimental Psychology: General 146*(7), 968.

Rabinowicz, W. (2002). Does practical deliberation crowd out self-prediction? *Erkenntnis 57*(1), 91–122.

Ramsey, F. P. (1929). *General propositions and causality*, pp. 145–63. Cambridge University Press.

Ramsey, F. P. (1931). Truth and probability. In *The Foundation of Mathematics and Other Logical Essays*. New York, Harcourt, Brace and Co.

Rollins, H. E. (2012). *The Letters of John Keats: Volume 1, 1814-1818: 1814-1821*, Volume 1. Cambridge University Press.

Russell, P. (2021). Hume on Free Will. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University.

Russell, S. and P. Norvig (1995). *Artificial intelligence: a modern approach*. Prentice-Hall, Inc.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Schick, F. (1979). Self-knowledge, uncertainty, and choice. *The British Journal for the Philosophy of Science 30*(3), 235–252.

Sellars, W. (1963). Philosophy and the scientific image of man. *Science, perception and reality 2*, 35–78.

Shackford, M. H. (1925). Hyperion. *Studies in Philology 22*(1), 48–60.

Simon, H. A. (1957). *Models of man; social and rational*. Wiley.

Skyrms, B. (1984). *Pragmatics and Empiricism*. Yale University Press.

Skyrms, B. (1987). Dynamic coherence and probability kinematics. *Philosophy of Science 54*(1), 1–20.

Skyrms, B. (1990). *The dynamics of rational deliberation*. Harvard University Press.

Skyrms, B. (1994). Darwin meets the logic of decision: Correlation in evolutionary game theory. *Philosophy of Science 61*(4), 503–528.

Sneed, J. D. (1966). Strategy and the logic of decision. *Synthese*, 270–283.

Spohn, W. (1977). Where luce and krantz do really generalize savage's decision model. *Erkenntnis*, 113–134.

Spohn, W. (2007). The core of free will. *Thinking About Causes. From Greek Philosophy to Modern Physics*, 297–309.

Steele, K. and H. O. Stefánsson (2020). Decision Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University.

Sterkenburg, T. F. (2016). Solomonoff prediction and occam's razor. *Philosophy of Science 83*(4), 459–479.

Thorpe, C. D. (1935). *Complete Poems and Selected Letters*. Odyssey Press.

Van Fraassen, B. C. (1977). Relative frequencies. *Synthese*, 133–166.

Van Fraassen, B. C. (1980). Rational belief and probability kinematics. *Philosophy of Science 47*(2), 165–187.

Vavova, K. (2016). Deliberation and prediction: It's complicated. *Episteme 13*(4), 529–538.

Velleman, J. D. (1989). Epistemic freedom. *Pacific Philosophical Quarterly 70*(1).

Villegas, C. (1964). On qualitative probability /*sigma*-algebras. *The Annals of Mathematical Statistics 35*(4), 1787–1796.

Von Neumann, J. and O. Morgenstern (1953). *Theory of games and economic behavior*. Princeton university press.

Wallace, D. (2012). *The emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford University Press.

Weisberg, J. (2011). Varieties of bayesianism. *Inductive logic 10*, 477–551.