

# Consistency in Choice and Credence

ARCHIVES

by

Brian Hedden

A.B. Philosophy  
Princeton University, 2006

Submitted to the Department of Linguistics and Philosophy in Partial  
Fulfillment of the Requirements for the Degree of

Doctor of Philosophy


at the

Massachusetts Institute of Technology  
September 2012


©Brian Hedden. All Rights Reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.


Signature of Author:

  
Department of Linguistics and Philosophy  
May 21, 2012

Certified By:

  
Caspar Hare  
Associate Professor of Philosophy  
Thesis Supervisor

Accepted By:

  
Roger White  
Associate Professor of Philosophy  
Acting Head of Graduate Studies

# Consistency in Choice and Credence

by

Brian Hedden

Submitted to the Department of Linguistics and Philosophy  
on May 21, 2012 in Partial Fulfillment of the  
Requirements for the Degree of Doctor of Philosophy in  
Philosophy

## Abstract:

This thesis concerns epistemic and practical rationality. That is, it is about what to believe and what to do. In Chapter 1, I argue that theories of practical rationality should be understood as evaluating *decisions* as opposed to ordinary sorts of non-mental actions. In Chapter 2, I use the machinery developed in Chapter 1 to rebut ‘Money Pump’ or ‘Diachronic Dutch Book’ arguments, which draw conclusions about rational beliefs and preferences from premises about how rational agents will behave over time. In Chapter 3, I develop a new objection to the Synchronic Dutch Book Argument, which concludes that rational agents must have probabilistic degrees of belief in order to avoid predictable exploitation in betting scenarios.

Thesis Supervisor: Caspar Hare

Title: Associate Professor of Philosophy

## Acknowledgements

Many people were of great help in developing the ideas contained in this dissertation. I would like to Ryan Doody, Dan Greco, Richard Holton, Heather Logue, Vann McGee, Tyler Paytas, Agustín Rayo, Miriam Schoenfeld, Paulina Sliwa, Matthew Noah Smith, and Steve Yablo, as well as audiences at the 2011 MITing of the Minds Conference, the MIT MATTI reading group, the MIT Epistemology Reading Group, the MIT Work in Progress seminar, the 2011 Bellingham Summer Philosophy, and the 2011 Rocky Mountain Ethics Congress, for very helpful discussion.

I am especially grateful to the members of my dissertation committee, who were invaluable throughout my graduate studies. This dissertation could not have been written without the help of Robert Stalnaker, Roger White, and Caspar Hare, who chaired the dissertation committee.

# Contents

<b>1</b>	<b>Introduction</b>	<b>6</b>
<b>2</b>	<b>Options and the Subjective <i>Ought</i></b>	<b>9</b>
2.1	Introduction . . . . .	9
2.2	The Subjective <i>Ought</i> . . . . .	10
2.3	The Problem of Options . . . . .	12
2.4	First Pass: Options as Actual Abilities . . . . .	13
2.5	Second Pass: Options as Believed Abilities . . . . .	18
2.6	Third Pass: Options as Known Abilities . . . . .	19
2.7	Options as Decisions . . . . .	20
2.8	Costs of Decisions . . . . .	24
2.9	Chisholm's Paradox . . . . .	25
2.10	Conclusion . . . . .	30
<b>3</b>	<b>Options and Diachronic Tragedy</b>	<b>33</b>
3.1	Introduction . . . . .	33
3.2	Diachronic Tragedy . . . . .	34
3.2.1	Preference Shifts . . . . .	34
3.2.2	Time Bias . . . . .	35
3.2.3	Intransitive Preferences . . . . .	37
3.2.4	Imprecise Preferences . . . . .	37
3.2.5	Infinite Decisions . . . . .	38
3.2.6	Formal Epistemology . . . . .	39
3.2.7	Common Structure . . . . .	40
3.3	The No Way Out Argument . . . . .	42
3.4	Options as Decisions . . . . .	45
3.5	Decisions and Diachronic Tragedy . . . . .	50
3.6	Is Everything then Permitted? . . . . .	54

3.7	Conclusion . . . . .	56
<b>4</b>	<b>Incoherence Without Exploitability</b>	<b>57</b>
4.1	Introduction . . . . .	57
4.2	The Canonical Dutch Book Argument . . . . .	59
4.3	How Credences License Actions . . . . .	61
4.4	Negation Coherence and the Limits of the DBA . . . . .	63
4.5	Incoherence without Exploitability . . . . .	64
4.6	An Objection: Credences Constitutively Linked to Fair Betting Quotients? . . . . .	67
4.7	Where do We Go From Here? . . . . .	68
	4.7.1 Impossible to Violate Negation Coherence? . . . . .	68
	4.7.2 Irrational to Violate Negation Coherence? . . . . .	70
4.8	Conclusion . . . . .	73

# Chapter 1

## Introduction

This dissertation is about what I ought to believe and what I ought to do. That is, it concerns epistemic and practical rationality. It centers on three themes: first, the applicability of formal models of rationality to the sorts of situations typically encountered by finite agents like us living in a messy, complex world; second, whether the requirements of rationality apply to temporally extended agents, or instead only to time-slices thereof; and third, the interaction between epistemic rationality and practical rationality.

Philosophers, economists, and mathematicians have developed fruitful and elegant formal models of epistemic and practical rationality. By far the most prominent such theory is Bayesianism. In the practical case, in order to yield a verdict about what one ought to do, the Bayesian framework requires three things as inputs: probabilities, utilities, and a set of options. Bayesianism is so easily applied in idealized cases like betting on Roulette precisely because it is clear what the relevant probabilities, utilities, and options are - the probabilities are given by the physical symmetries of the wheel, the utilities are tied to possible monetary payoffs, and the options are all the bets allowed by the rules of the game. But in more mundane decision situations outside Las Vegas, what are the probabilities, utilities, and options that we should plug into the Bayesian machinery? Unless we can answer this question, Bayesianism will be inapplicable to daily life.

In Chapter 1 ('Options and Subjective *Ought*'), I confront the problem of saying what your options are in a decision situation. (Probabilities and utilities have already been widely discussed.) I argue that we must conceive of your options as consisting of all and only the *decisions* you are presently able to make. Thus, the subjective *ought* applies only to mental acts, and not

to the non-mental acts that philosophers often evaluate for rational permissibility. In this way, Chapter 1 constitutes a partial defense of the applicability of Bayesian tools to everyday life, as it shows that one obstacle to applying Bayesian tools in ordinary circumstances can be overcome.

This theory of options is time-slice centric, in the sense that it holds that the *ought* of practical rationality applies only to instantaneous or near-instantaneous actions (acts of decision-making) that can be performed by a particular time-slice of an agent. More generally, I am sympathetic to the view that all requirements of rationality, whether epistemic or practical, are time-slice centric. Requirements of rationality apply only to time-slices, and not to temporally extended ‘space-time worms’ (mereological sums of time-slices). The motivation for this time-slice centric view of rationality is an internalist one, on which what you ought to believe or do depends only on your present mental state, and not on things that could vary independently of your mental state. For this reason, I think that requirements of rationality must depend only on your present mental state and must be requirements that can be fulfilled by your present mental state, independent of what your past or future time-slices may do.

In Chapter 2 (‘Options and Diachronic Tragedy’), I use this time-slice centric view of options to undermine Diachronic Dutch Book Arguments. Philosophers have used these arguments to support a wide range of requirements of rationality (e.g. that you ought to defer to the beliefs that you believe you will later have), on the grounds that violating that requirement will lead you to act over time in a manner that is to your own acknowledged disadvantage. Diachronic Dutch Book Arguments rest on the assumption that sequences of actions can be evaluated for rationality. But according to my time-slice centric account of options, this assumption is false. If your options consist of only the decisions your present time-slice can make, then sequences of actions are not evaluable for rationality, and so the fact that (e.g.) being time-biased requires you to act over time to your own disadvantage does not show that it is irrational to be time-biased.

In Chapter 3 (‘Incoherence without Exploitability’), I turn my attention to the *Synchronic* Dutch Book Argument, and attempt to show that it is likewise unsound. The Synchronic Dutch Book Argument aims to show that your degrees of belief are irrational unless they conform to the axioms of the probability calculus. It holds that violating the probability calculus will lead you to be willing to accept each member of a set of bets which together guarantee you a financial loss (which is purportedly irrational). But I argue

that on a careful treatment of the relationship between degrees of belief and actions, this is false. It is possible to have degrees of belief that violate the probability calculus without being committed to this sort of irrational betting behavior.

Both the Diachronic and Synchronic Dutch Books Arguments draw conclusions about epistemic rationality (e.g. that you ought to defer to your anticipated future beliefs; that your degrees of belief ought to conform to the probability calculus) from premises about practical rationality (that you ought not perform predictably disadvantageous actions or courses of actions). While I do not have any objection in principle to inferring facts about epistemic rationality from facts about practical rationality or vice versa, I hope to have shown at a minimum that we must tread lightly when doing so. Moreover, if I have been successful in rebutting the two most prominent sorts of argument that relate epistemic rationality to practical rationality, perhaps epistemic and practical rationality are further apart than they are often thought to be.

In sum, we want our theories of rationality, including formal ones like Bayesianism, to be applicable to the circumstances in which we find ourselves. One challenge to applying theories of rationality to these messy, non-idealized circumstances is coming up with a theory of what your options are in a given decision situation. I argue for a time-slice centric view of options, on which your options at a particular time are the decisions that can be made by your present time-slice. This time-slice centric view of rationality leads to a rejection of Diachronic Dutch Book Arguments. And this, together with a separate rebuttal of the Synchronic Dutch Book Argument, constitutes an undermining of some of the most prominent attempts to connect epistemic rationality to practical rationality.



## Chapter 2

# Options and the Subjective *Ought*

### 2.1 Introduction

Determining what you ought to do can be broken down into two stages. The first stage is determining what your options are, and the second stage is ranking those options. While the second stage has been widely explored by philosophers of all stripes, from ethicists to decision theorists to epistemologists to action theorists, the first stage has gone largely unaddressed. And yet, without a theory of how to conceive of your options, the theory of practical rationality - of how you ought to act - will be incomplete.

I will argue that the fact that what you ought to do depends on your uncertainty about the world ultimately forces us to conceive of your options as consisting of all and only the *decisions* you are presently able to make. In this way, *oughts* apply only to decisions, and not to the non-mental acts that we ordinarily evaluate for moral and rational permissibility.

This change in our conception of your options is not of mere bookkeeping interest; it has substantive implications for first-order normative theorizing. First, it directly takes into account the potential costs and benefits of the decision itself in determining what you ought to do. Second, it provides a principled solution to Chisholm's Paradox, in which your doubts about your own self-control seem to give rise to conflicting claims about what you ought to do. These two cases provide direct support for my theory of what your options are, supplementing the more theoretical reasons considered earlier.

## 2.2 The Subjective *Ought*

In this paper, I will be focused on the sense of *ought* in which what you ought to do depends on your beliefs about how the world is. Consider: Your friend has a headache, and you have some pills that you justifiably believe to be pain relievers. But you're wrong. They are really rat poison. Ought you give the pills to your friend?

While there may be a sense in which the answer is 'no,' there is also a sense in which the answer is 'yes.' Call the sense of *ought* in which you ought to give your friend the pills the *subjective ought*. What you subjectively ought to do depends not on how the world actually is, but on how you believe the world to be. Since you believe the pills to be pain relievers, you subjectively ought to hand them over, even though your belief is false.<sup>1</sup>

The subjective *ought* has three important roles to play. First, the subjective *ought* is supposed to give you guidance about how to proceed (morally or prudentially speaking), given your uncertainty about the world. It is to be action-guiding, in at least the minimal sense that you are typically in a position to know what you subjectively ought to do.<sup>2</sup>

Second, the subjective *ought* plays an evaluative role. In ethics, it links up tightly with praise and blame. If you give your friend the pills, you are praiseworthy, or at least exempt from blame, for having done so, despite the disastrous consequences, since you did what you subjectively ought to have done. With respect to prudential rationality, you are subject to rational

---

<sup>1</sup>The sense of *ought* in which you ought not give your friend the pills is often called the *objective ought*. In the case of prudential rationality, what you objectively ought to do is whatever would in fact maximize your utility, while what you subjectively ought to do is bring about whichever proposition has highest *expected* utility. In ethics, consequentialists will likely say that what you objectively ought to do is whatever would maximize moral value (total world happiness, say), while what you subjectively ought to do is bring about whichever proposition has highest *expected* moral value. The objective/subjective distinction can also be drawn in non-consequentialist moral theories, although there is less consensus on how exactly to do so.

<sup>2</sup>To emphasize - I am understanding the requirement that the subjective *ought* be 'action-guiding' as the requirement that you be *in a position to know* what you ought to do. Thus, for the subjective *ought* to be action-guiding, it is not required that you always in fact know what you ought to do (for you might make a mistake or fail to consider the question), nor is it required that you consciously employ the theory of the subjective *ought* in coming to a conclusion about what you ought to do. All that is required for the subjective *ought* to be action-guiding, in my sense, is for facts about what you ought to do to be in principle accessible to you.

criticism unless you do what you prudentially subjectively ought to do.

Third, in the case of prudential rationality, the subjective *ought* plays a role in the prediction and explanation of behavior. Knowing what you believe and desire, we can predict what you will do, against the background assumption that you are rational. We predict that you will do that which you (prudentially) subjectively ought to do. Why is this? It is close to an analytic truth that fully rational agents successfully fulfill their rational requirements; they do what they rationally ought to do (that is, what they prudentially subjectively ought to do), believe what they rationally ought to believe, and so forth. So, given the assumption that you are rational, and given that this entails that you fulfill your rational requirements, it follows straightforwardly that you will perform the action that you subjectively ought to perform.<sup>3</sup> And we can explain why you did something by pointing out that you are rational and that, given your beliefs and desires, that was the thing you subjectively ought to have done.

So, the subjective *ought* should play an action-guiding, evaluative, and predictive/explanatory role in our normative theorizing. I favor this view, on which one sense of *ought* plays all three roles, both because it is parsimonious and because the three roles are interrelated. To take just one example, it is plausible that a sense of *ought* cannot play the evaluative role without also playing the action-guiding role (and vice-versa), since an agent cannot be criticized for performing some action or having some belief if she was in no way even in a position to know that she oughtn't have performed that action or held that belief.<sup>4</sup> Importantly, however, my argument in this paper does not depend on this assumption that there is one sense of *ought* which can play all three roles. This is because each of the desiderata presented below that support my favored view of options is motivated by each

---

<sup>3</sup>Importantly, we only predict that you will do what you subjectively ought to do when we hold onto the background assumption that you are rational. But often, we have evidence that you fall short of ideal rationality in various respects, and in these cases we will not want to predict that you will do what you subjectively ought to do. For instance, we may have evidence from behavioral economics that you employ certain biases and heuristics that lead you to be irrational in certain systematic ways, and if such biases and heuristics are relevant in the case at hand, we will not want to predict that you will in fact do what you subjectively ought to do.

<sup>4</sup>This is just to express sympathy with internalism about practical and epistemic rationality. Externalists will predictably be unsympathetic, but in some sense this paper can be seen as exploring the viability of internalism about practical rationality and determining how internalists should conceive of a decision-maker's options.

of these three roles. Therefore, one will still be pushed toward my view of options even if one thinks that we will need multiple *oughts*, one of which will play the action-guiding role, another the evaluative role, and a third the predictive/explanatory role.<sup>5</sup>

## 2.3 The Problem of Options

Giving a theory of the subjective *ought* requires giving a theory of what your options are. Because the subjective *ought* is sensitive to your uncertainty about the world, this theory of options must also take into account this uncertainty if the subjective *ought* is to play its action-guiding, evaluative, and predictive/explanatory roles.

The problem of specifying what your options are, in a way that is appropriately sensitive to your uncertainty about the world, can be made precise by considering expected utility theory, the dominant account of the subjective *ought* of prudential rationality. (For clarity, I focus on prudential rationality in this paper, though the considerations I raise will apply equally to the case of ethics.) Expected utility theory provides a framework for assigning numbers to propositions, relative to a credence function  $P$  (representing the agent's doxastic, or belief-like, state) and a utility function  $U$  (representing the agent's conative, or desire-like, state). The expected utility of a proposition  $A$  is the sum of the utilities assigned to the possible outcomes  $O_i$ , weighted by the agent's credence that  $O_i$  will come about, conditional on  $A$ .<sup>6</sup> More formally:

$$\textbf{Expected Utility: } EU(A) = \sum_i P(O_i|A)U(O_i)$$

Expected utility theory, then, provides a way of ranking propositions. The connection between this ranking and prudential rationality is standardly expressed in the slogan, 'You ought to maximize expected utility.' That is, you ought to bring about the proposition with highest expected utility.

But now a problem arises. Expected utilities can be assigned to any proposition whatsoever. Consider, for example, the proposition that someone

---

<sup>5</sup>Thanks to Matthew Noah Smith for raising this worry.

<sup>6</sup>This is the definition of expected utility employed in *Evidential Decision Theory* and is sometimes called *evidential expected utility*. The definition of expected utility employed in *Causal Decision Theory* is slightly more complex, but the distinction between evidentialist and causalist views of expected utility will not matter in what follows.

discovered a cure for cancer two weeks ago. This proposition has a very high expected utility. But even if this proposition has higher expected utility than any other proposition, there is no sense in which I ought to bring it about that someone discovered a cure for cancer two weeks ago.<sup>7</sup> Intuitively, the expected utility assigned to the proposition that someone cured cancer two weeks ago is irrelevant to the question of what I ought to do now because bringing about this proposition simply isn't one of my options!

Therefore, in order to have a theory of what I ought to do, we need some way of specifying a narrower set of propositions, such that what I ought to do is bring about the proposition with highest expected utility *in that narrower set*. Let us call such a narrower set of propositions a *set of options*. Our task, then, is to say what counts as a set of options.

In what follows, I argue that a theory of options must satisfy three desiderata: First, if something is an option for you, you must be able to do it. Second, if something is an option for you, you must *believe* that you are able to do it. Third, what your options are must supervene on your beliefs and desires. I reject three initially tempting theories of options on the grounds that they each violates at least one of these desiderata. I then present my own theory of options, which I argue does satisfy the desiderata: your options are all and only the *decisions* you are presently able to make.

## 2.4 First Pass: Options as Actual Abilities

Consider again the example of the proposition that someone found a cure for cancer a couple weeks ago. One obvious reason for thinking that there's no sense in which I ought to bring it about that someone has found a cure a couple weeks ago is that *that's just not something I can do!* We must respect the principle that *ought implies can* if the subjective *ought* is to play the action-guiding, evaluative, and predictive roles that are supposed to be played by the subjective *ought*. First, the theory gives poor guidance to an agent if it tells her to do something that she cannot do. Second, an agent is not subject to any rational criticism if she fails to do something which in fact she couldn't have done. Third, we would not want to predict that a rational

---

<sup>7</sup>Of course, there would certainly be other propositions with higher expected utility, such as the proposition that someone discovered a cure for cancer *and* deposited \$10,000 in my bank account. In fact, it may be that there is no proposition with highest expected utility.

agent would do something which in fact she cannot do. This yields a first desideratum for a theory of options:

**Desideratum 1:** If a proposition P is a member of a set of options for an agent S, then S is able to bring about P.

A minimal theory, on which *ought implies can* is the only restriction on what counts as an option, would be the following:

**Proposal 1:** A set of propositions is a set of options iff it is a maximal set of mutually exclusive propositions, each of which is such that the agent has the ability to bring it about.<sup>8</sup>

Not only is Proposal 1 intuitively attractive, but it also has a formidable pedigree, having been defended by many prominent decision theorists, including Richard Jeffrey and David Lewis.<sup>9</sup> But Proposal 1 is too permissive about what can count as an option for an agent.

First, Proposal 1 can yield unacceptable results in cases where the agent is uncertain or in error about her own abilities. Consider:

**Raging Creek:** Jane is hiking along the trail when she comes to a raging creek. She is in fact able to ford the creek (where this entails getting across), head upstream to look for an easier crossing, or turn back. Among these three things that she is able to do, fording the creek has highest expected utility, since she

---

<sup>8</sup>The set must be maximal in the sense that there is no other proposition incompatible with the members of that set which is also such that the agent has the ability to bring it about. Note that this proposal allows for the possibility of multiple sets of options for an agent, since we can cut up the things that she is able to bring about in more or less fine-grained ways and still have a maximal set of mutually exclusive propositions, each of which she is able to bring about.

<sup>9</sup>Jeffrey (1965, 84) regards options as acts, where ‘An act is then a proposition which is within the agent’s power to make true if he pleases.’ And in ‘Preference among Preferences,’ he writes that ‘To a first approximation, an option is a sentence which the agent can be sure is true, if he wishes it to be true’ (Jeffrey (1992, 164)). In ‘Causal Decision Theory,’ Lewis writes, ‘Suppose we have a partition of propositions that distinguish worlds where the agent acts differently...Further, he can act at will so as to make any one of these propositions hold; but he cannot act at will to make any proposition hold that implies but is not implied by (is properly included in) a proposition in the partition. The partition gives the most detailed specifications of his present action over which he has control. Then this is a partition of the agents’ alternative *options*’ (Lewis (1981, 7)).

would then be able to continue hiking with no detour. But she has serious doubts about whether she can in fact ford the creek. After all, the water is up to mid-thigh and flowing fast.

Ought Jane ford the creek? I suggest that the answer is ‘no.’ First, the subjective *ought* is supposed to be action-guiding, in that what an agent subjectively ought to do depends solely on facts that the agent is in a position to know. But since Jane is not in a position to know that she is able to ford the creek, she is not in a position to know that she ought to ford the creek. Second, Jane is not subject to any rational criticism for failing to ford the creek. It would be bizarre to call her irrational for failing to ford it, given that she had serious doubts about her ability to do so. Third, we would not predict that Jane will ford the creek, given the information that she is rational and has these beliefs and desires. This case suggests that we must add a second desideratum for a theory of options which requires that something can count as an option for an agent only if she believes she is able to to it:

**Desideratum 2:** If a proposition P is a member of a set of options for an agent S, then S believes that she is able to bring about P.<sup>10</sup>

Another way to see the need for Desideratum 2 is this: The calculation of the expected utility of an action  $\phi$  fails to take into account the possibility that the agent might try to  $\phi$  but fail to succeed in this effort. This is because the expected utility of  $\phi$  is the sum the values of the possible outcome states, weighted by probability of that outcome state *given the assumption that the act  $\phi$  is performed*. But the probability of an outcome in which the agent tries to  $\phi$  but fails, conditional on her  $\phi$ -ing, is of course zero! So the disutility

---

<sup>10</sup>One might prefer, here and elsewhere, to replace talk of what the agent *actually* believes and desires with talk of what the agent *ought* to believe and desire. In this way, what an agent subjectively ought to do would depend not on what she believes and desires, but on what she ought to believe and desire. Importantly, adopting this view will not affect the arguments in this paper; one will still be pushed to adopt my favored theory of options. I will continue to put things in terms of the agent’s actual beliefs and desires for the sake of consistency, and also because I favor keeping epistemic and practical rationality distinct. In cases where an agent has beliefs (or desires) that she ought not have, but acts in a way that makes sense, given those misguided beliefs (or desires), we should criticize her for being epistemically irrational without also accusing her of practical irrationality.

of any outcome in which the agent tries to  $\phi$  but fails is multiplied by zero in the expected utility calculation, and hence not taken into account in the evaluation of the act of  $\phi$ -ing. But as **Raging Creek** illustrates, what would happen if the agent were to try but fail to  $\phi$  is often very, very important to the agent's own deliberations and to our evaluation of her! Our response to this problem should be to think of an agent's options as including only actions that the agent is confident she can perform and thereby exclude from consideration as options any actions which the agent thinks she might try but fail to perform. Hence Desideratum 2.<sup>11</sup>

Second, since actual abilities do not supervene on beliefs and desires, Proposal 1 entails that what an agent subjectively ought to do will not supervene on beliefs and desires, either. If two agents have the same beliefs and desires but different abilities, then the sets of options relevant for assessing what they ought to do will also be different, with the result that they ought to do quite different things.<sup>12</sup> Consider:

**Jane's Doppelgänger:** Jane, facing the raging creek, is in fact able to ford it, and among the things she is able to do, fording the creek is best (has highest expected utility). But her doppelgänger Twin Jane, who has the same beliefs and desires as Jane, faces her own raging creek, but Twin Jane is unable to ford it. Among the things that Twin Jane is able to do, turning back and heading home is best (has highest expected utility).

On Proposal 1, Jane ought to ford the creek, while Twin Jane ought to turn back and head home. But this is an implausible result. First, if Jane

---

<sup>11</sup>Heather Logue has pointed out to me that Desideratum 2 may not actually be necessary to motivate my favored theory of options, since my theory of options may also be the only one which can satisfy both Desideratum 1 and Desideratum 3 (below). Still, I include Desideratum 2 since I think it is a genuine desideratum, even if it is not needed to motivate my view.

<sup>12</sup>Of course, if we think of *oughts* as attaching to act tokens, rather than act types, there is a harmless sense in which what an agent ought to do will not supervene on that agent's mental states. Perhaps my physically identical doppelgänger and I are in exactly the same mental states, but while I ought to bring it about that *I* donate money to charity, my doppelgänger ought to bring it about that *he* donate money to charity. There is nothing disconcerting about this. It would be more problematic if what an agent ought to do, put in terms of act types like *donating to charity* (perhaps modelled using sets of centered worlds instead of sets of worlds), failed to supervene on beliefs and desires. This more worrying type of failure of supervenience is entailed by Proposal 1.



and Twin Jane are in exactly the same mental state but ought to do quite different things, neither is in a position to determine that she ought to ford the creek rather than turn back, or vice versa. So again, the subjective *ought* would be in an important sense insufficiently action-guiding.<sup>13</sup> Second, it is implausible to evaluate Jane and Twin Jane differently (criticizing the one but praising the other) if they perform the same action (heading home, say) after starting out in the same mental state, but this is what would be required if we say that *oughts* fail to supervene on beliefs and desires and that Jane and Twin Jane differ in what they ought to do. Third, consider the role of the subjective *ought* in the prediction of behavior. We predict that an agent will perform the action that she ought to perform. So we should predict that Jane will ford the creek, while Twin Jane will immediately do an about-face and head straight home. But this is bizarre! They are in exactly the same mental state, after all! If they displayed such radically different behavior, it would appear that at least one of them wasn't fully in control of her actions (and hence not fully rational). In general, the failure of what an agent ought to do to supervene on her beliefs and desires would entail a sort of externalism about practical rationality. But externalism about practical rationality is a tough pill to swallow, as it would keep the subjective *ought* from adequately playing its action-guiding, evaluative, and predictive roles in our theorizing. This suggests adding a third desideratum for a theory of options:

**Desideratum 3:** If something is an option for an agent S, then it is an option for any agent with the same beliefs and desires as S.

---

<sup>13</sup>Of course, supervenience of what an agent ought to do on her beliefs and desires is not by itself sufficient for her to be in a position to know what she ought to do. She must also know her beliefs and desires. The important point is that self-knowledge and supervenience of *oughts* on beliefs and desires are individually necessary and jointly sufficient for the agent to be in a position to know what she ought to do. Therefore, if supervenience fails, then even a self-knowing agent would not be in a position to know what she ought to do. Importantly, knowledge of one's beliefs and desires is already required for one to be in a position to know what one ought to do, since even knowing what one's options are, one needs to know what one believes and desires in order to know how to rank those options. Provided that an agent's options supervene on her beliefs and desires, there are no obstacles to her being in a position to know what her options are that are not already obstacles to her being in a position to know how those options are to be ranked.

While Proposal 1 satisfies Desideratum 1, it allows too many things to count as options for an agent and violates two other compelling desiderata.

## 2.5 Second Pass: Options as Believed Abilities

The subjective *ought* is supposed to be sensitive to how the agent takes the world to be. Now, as noted earlier, there are two stages to determining what an agent ought to do. The first stage involves identifying the agent's options, and the second stage involves ranking those options. At the second stage, we have a compelling framework (expected utility theory) for ranking options in light of an agent's beliefs and desires. But this progress goes to waste if at the first stage we characterize an agent's options in a manner determined not by how the agent believes the world to be, but only by how the world actually is. Proposal 1 made this mistake by taking an agent's options to be determined by her actual abilities, regardless of what she believed about her abilities.

With this in mind, we might conceive of an agent's options as all and only the things she believes she is able to do. More precisely:

**Proposal 2:** A set of propositions is a set of options iff it is a maximal set of mutually exclusive propositions, each of which is such that the agent believes she has the ability to bring it about.<sup>14</sup>

This proposal satisfies Desiderata 2 and 3, which were Proposal 1's downfall. First, in cases where an agent can do something, but doubts whether she can do it, that thing will not count as an option for her. Hence, when Jane arrives at the raging creek and doubts whether she can ford it (even though she in fact can), fording the creek will not be one of her options, and so we will avoid the counterintuitive result that she ought to ford the creek. Second, because an agent's options are wholly determined by her beliefs (in particular her beliefs about her own abilities), what an agent ought to do will supervene on her beliefs and desires, as required.

But Proposal 2 is also a step back, for it fails to satisfy Desideratum 1 (*ought implies can*). If Jane believes that she can ford the creek, and fording

---

<sup>14</sup>Again, the set must be 'maximal' in the sense that there is no other proposition incompatible with the members of that set which is also such that the agent has the ability to bring it about.

the creek has highest expected utility among the things that Jane believes she can do, then we get the result that Jane ought to ford the creek, even if she is in fact unable to do so. And again, *ought implies can* is non-negotiable. First, our theory gives an agent poor guidance if it tells her to do something that she in fact cannot do. Second, an agent is subject to no rational criticism (or blame, in the ethical case) for failing to do something she was unable to do. And third, we would not want to predict that an agent will perform an action that she is unable to perform.<sup>15</sup>

## 2.6 Third Pass: Options as Known Abilities

Proposal 1 failed because it had an agent's options being determined solely by how the world actually is, irrespective of how the agent believes the world to be. Proposal 2 failed because it had an agent's options being determined by how she believes the world to be, irrespective of how it actually is. Perhaps these problems can be solved by taking an agent's options to be the things that are deemed to be options by both Proposal 1 and Proposal 2. That is, we might take an agent's options to be the set of propositions that she correctly believes she can bring about. At this point, we might even replace the mention of true belief with reference to knowledge, giving us:

**Proposal 3:** A set of propositions is a set of options iff it is a maximal set of mutually exclusive propositions, each of which is such that the agent knows that she is able to bring it about.<sup>16</sup>

This conception of an agent's options satisfies Desiderata 1 and 2. It will yield the attractive result that if an agent ought to do something, then she is able to do it and believes that she is able to do it.

But this proposal violates Desideratum 3, that what an agent's options are should supervene on her beliefs and desires. This is because if one agent knows that she is able to perform some given action, while another falsely

---

<sup>15</sup>A close cousin of Proposal 2 would characterize an agent's options in normative terms, so that an agent's options consist not of the things which she actually believes she can do, but rather of the things which she *ought* to believe she can do. This proposal, however, will likewise violate Desideratum 1 and is unacceptable on this account.

<sup>16</sup>Once again, the set must be 'maximal' in the sense that there is no other proposition incompatible with the members of that set which is also such that the agent knows that she is able to bring it about.

believes that she is able to perform that action, then it will count as an option only for the first agent, even if the two of them have the same beliefs and desires. Recall **Jane's Doppelgänger**. Jane and Twin Jane have the same beliefs and desires. But Jane is able to ford the creek and knows that she is able to do so, whereas Twin Jane is unable to ford the creek and so falsely believes that she is able to do so. So on Proposal 3, fording the creek will count as an option for Jane but not for Twin Jane, and so it may be that Jane ought to ford the creek whereas Twin Jane ought to do something quite different. But as we saw earlier, this is an unacceptable result.

The root problem is this: In order for the subjective *ought* to play its theoretical roles, it is important that an agent is at least typically in a position to know what her options are. But Proposal 3 does not have this result. On Proposal 3, an agent will typically only be in a position to know, of the things that are options for her, that they are options for her<sup>17</sup>; she will not typically be able to know, of the things that are not options for her, that they are not options for her. For instance, if Twin Jane justifiably but falsely believes that she is able to ford the creek, then she will not be in a position to know that fording the creek is not an option for her. So, it is important that an agent typically be able to tell whether or not something is an option for her, but Proposal 3 fails to give this result because it violates Desideratum 3, that an agent's options should supervene on her beliefs and desires.

## 2.7 Options as Decisions

The proposals considered above each failed to satisfy one or more of our three desiderata on a theory of an agent's options. First, if P is an option for an agent, then she is actually able to bring about P. Second, if P is an option for an agent, then she believes she is able to bring about P. Third, if P is an option for an agent, then it is also an option for any agent with the same beliefs and desires.

But is it even *possible* for an account to satisfy all three of these desiderata? I think so. The key is to think of an agent's options as consisting

---

<sup>17</sup>Actually, this assumes a perhaps controversial application of the KK principle, which states that when an agent know that P, she is in a position to know that she knows P. This is because on Proposal 3, knowing that something is an option for you requires knowing that you know you are able to bring it about. If KK fails, then so much the worse for Proposal 3.

of the *decisions* open to her. Indeed, in cases where an agent is uncertain about whether she is able to do something like ford a creek, it is natural to think that the option we should really be evaluating is her *deciding* to ford the creek, or deciding to try to ford the creek. I suggest that the things that are in the first instance evaluated for rationality or irrationality are *always* the agent's decisions. This gives us a new proposal, which I will call **Options-as-Decisions**:

**Options-as-Decisions:** A set of propositions is a set of options for agent  $S$  at time  $t$  iff it is a maximal set of mutually exclusive propositions of the form  $S$  *decides at  $t$  to  $\phi$* , each of which  $S$  is able to bring about.<sup>18</sup>

Then, the highest-ranked such proposition (by expected utility) will be such that the agent ought to bring it about. That is, she ought to make that decision. (I am intending 'S decides to  $\phi$ ' to be read in such a way that is incompatible with 'S decides to  $\phi \wedge \psi$ ,' even though there is a sense in which if you decide to do two things, you thereby decide to do each one. If you have trouble getting the intended reading, just add in 'only' to the proposal, so that a set of options is a maximal set of propositions of the form  $S$  *only decides at  $t$  to  $\phi$* , each of which the agent is able to bring about.)

But will **Options-as-Decisions** satisfy the three desiderata? This all depends on what it takes for an agent to be able to make a given decision. I do not want to commit myself to any particular account of what it takes to be able to make a decision, but I will argue that on any attractive theory of decisions, which decisions an agent is able to make will supervene on her beliefs and desires and thereby allow **Options-as-Decisions** to satisfy the three desiderata.

As one example of a prominent theory of decisions, Bratman (1987) holds that you can make any decision that you do not believe would be ineffective: You are able to decide to  $\phi$  if and only if you do not believe that, were you

---

<sup>18</sup>Once again, 'maximal' means that there is no proposition of the form  $S$  *decides at  $t$  to  $\phi$*  which is not a member of the set but which is incompatible with each member of the set. Note that maximality and mutual exclusivity apply not to the *contents* of decisions, but to propositions about which decision was made. Hence the set  $\{S$  *decides at  $t$  to  $\phi$* ,  $S$  *decides at  $t$  not to  $\phi$*  $\}$  will not count as a set of options, since it does not include propositions about other decisions that  $S$  might have made (e.g. the proposition that  $S$  decides at  $t$  to  $\psi$ ).

to decide to  $\phi$ , you would fail to  $\phi$ .<sup>19</sup>

If Bratman is right, then **Options-as-Decisions** satisfies Desiderata 1-3. First, it satisfies Desideratum 1 (that if something is an option for an agent, then she is able to bring it about), since it is part of **Options-as-Decisions** that a proposition of the form *S decides at t to  $\phi$*  can count as an option for S only if S is able to bring about that proposition. Second, it satisfies Desideratum 2 (that if something is an option for an agent, then she believes she is able to bring it about), at least insofar as the agent knows what she believes. If an agent knows what her beliefs are, then she will know whether she believes that, were she to decide to  $\phi$ , she would  $\phi$ . And hence, she will know whether she is able to decide to  $\phi$ . Third, it satisfies Desideratum 3 (that what an agent's options are supervenes on her beliefs and desires), since on Bratman's view, which decisions an agent is able to make, and so what her options are, is entirely determined by her beliefs.

Even if Bratman's view is incorrect, other attractive views of abilities to make decisions still yield the result that **Options-as-Decisions** satisfies Desiderata 1-3. One might hold that in order to be able to decide to  $\phi$ , you must not only lack the belief that your decision to  $\phi$  would be ineffective (as Bratman holds); you must also have the belief that your decision to  $\phi$  would be effective.<sup>20</sup> Or one might hold that whether or not you are able to

---

<sup>19</sup>Actually, Bratman is discussing intentions, but I think that the relevant considerations apply equally to decisions, insofar as there is any difference between decisions and intentions. This theory of abilities to make decisions gains support from Kavka's Toxin Puzzle (Kavka (1983)). Suppose that in one hour, you will be offered a drink containing a toxin which will make you temporarily ill. Now, you are offered a large sum of money if you make the decision to drink the beverage. You will receive the money even if you do not then go ahead and drink the beverage; the payment depends only on your now making the decision to drink it. It seems that you cannot win the money in this case; you cannot decide to drink the beverage. Why not? Because you believe that, if you were to make the decision to drink the beverage, you would later reconsider and refuse to drink it. You cannot make a decision if you believe you will not carry it out. Supposing that this is the only restriction on agents' abilities to make decisions, we get Bratman's theory of abilities to make decisions.

<sup>20</sup>Anscombe (1957b) famously holds that in order to be able to decide to  $\phi$ , you do not even need to lack the belief that your decision to  $\phi$  would be ineffective. You can make decisions that you believe you will not carry out. For instance, as you are being led to the interrogation room, you can decide not to give up your comrades, even though you know you will crack under the torture. Some have interpreted Anscombe as holding that there are *no* restrictions on which decisions you are able to make. If this (admittedly somewhat implausible) view is true, **Options-as-Decisions** will still satisfy Desiderata

decide to  $\phi$  depends not just on your beliefs, but also on your desires. So, it might be that you are unable to decide to commit some horrible murder, even though you believe that, were you to manage to make this decision, you would indeed carry it out. You just have an incredibly deep aversion to making this decision, and this aversion prevents you from being able to do so.

But even if Bratman is wrong and one of these alternative accounts of abilities to make decisions is correct, **Options-as-Decisions** would still satisfy Desiderata 1-3, since even on these alternative accounts, which decisions you are able to make is wholly determined by your beliefs and desires. First, a decision would count as an option for an agent only if she is able to make it (by the wording of **Options-as-Decisions**), thereby satisfying Desideratum 1. Second, insofar as the agent knows what her beliefs and desires are, she will know which decisions she is able to make, thereby satisfying Desideratum 2. Third, since which decisions an agent is able to make will be fully determined by her beliefs and desires, which options she faces will supervene on her beliefs and desires, thereby satisfying Desideratum 3. So, in general, if which decisions you are able to make depends solely on your mental states, **Options-as-Decisions** will satisfy Desiderata 1-3.

But what if an agent's abilities to make decisions are restricted not just by her own mental states, but also by external forces? Frankfurt (1969) considers the possibility of a demon who can detect what's going on in your brain and will strike you down if he finds out that you are about to make the decision to  $\phi$ . Plausibly, you lack the ability to decide to  $\phi$ , even if you believe that, were you to decide to  $\phi$ , you would  $\phi$ . The possibility of such demons threatens the claim that which decisions you are able to make supervenes on your mental states, since which decisions you can make depends also on whether or not such a demon is monitoring you. It also threatens the claim that you are always in a position to know which decisions you are able to make, since you are not always in a position to know whether such a demon is watching you.

In response to this worry, I find it plausible that if a Frankfurtian demon is monitoring you with an eye toward preventing you from deciding to  $\phi$ , then you lack the capacity to exercise your rational capacities which is necessary in order for you to be subject to the demands of prudential rationality in

---

1-3, for trivially an agent will always be able to know which decisions she make make, and which decisions she can make will supervene on her beliefs and desires.

the first place. Suppose that the decision to  $\phi$  looks best out of all the decisions you believe you are able to make, but a demon will strike you down if it detects that you are about to  $\phi$ . What ought you to do in this case? Certainly, it is not that you ought to make some decision other than the decision to  $\phi$ , since all such decisions look inferior. And it is not the case that you ought to decide to  $\phi$ , since *ought* implies *can*. Instead, there simply isn't anything that you ought to *do*; rather, you ought to be in a state of being about to decide to  $\phi$ , where this will lead to your being struck down before you are actually able to *do* anything at all. The rational *ought* thus only applies to agents who are not being disrupted by Frankfurtian demons in this way, and so once we restrict our attention to agents to whom the rational *ought* applies, which options an agent has will both supervene on her beliefs and desires and be knowable by her.

I conclude that **Options-as-Decisions** will satisfy the three desiderata to which a theory of options is subject. Indeed, I think that it is the *only* theory of options which can satisfy these desiderata.

## 2.8 Costs of Decisions

The question of what your options are is not a mere terminological issue. In this section and the next, I consider two sorts of cases in which **Options-as-Decisions** yields attractive treatments of particular sorts of decision situations.

First, **Options-as-Decisions**, unlike other theories of options, directly takes into account the potential costs and benefits of the decision itself. Consider:

**God and Church:** You are deliberating about whether or not to skip Church tomorrow. You would prefer to stay home, but you believe that you will incur God's wrath if you decide to do so. However, you believe that God only punishes people who *decide* to avoid church; He does not punish forgetfulness or sleeping through alarms.

On **Options-as-Decisions**, God's punishment will be directly factored into the advisability of deciding to stay home. It is the decision itself which is evaluated for expected utility. Since you believe that this option (deciding



to stay home) will put you at the mercy of God's wrath, it has very low expected utility, and so you ought not decide to stay home.

But on Proposals 1-3, we can directly evaluate the options of staying home and attending church (rather than just decisions to do such things). Most likely, you have some high credence that if you stay home, it will be the result of a decision to stay home. But you also have some credence that if you stay home, it will be the result of forgetfulness or an ineffective alarm clock. In calculating the expected utility of staying home, the disutility of God's wrath is weighted by your credence that your staying home would follow a decision to do so. Then, if your credence that your staying home would be the result of a decision to do so (rather than forgetfulness or sleepiness) is sufficiently low, Proposals 1-3 will say that you ought to stay home, even in a case where **Options-as-Decisions** says that you ought to decide to attend church. On this point, I think that **Options-as-Decisions** has things right. If some action only looks good on the assumption that it will not be the result of a decision to perform that action, it seems mistaken to go on to say that you ought to perform that action. Proposals 1-3 sometimes have this unappealing result, while **Options-as-Decisions** never does.

## 2.9 Chisholm's Paradox

My account of options provides an attractive and well-motivated response to Chisholm's Paradox. Consider:

**Professor Procrastinate:**<sup>21</sup> You have been invited to write an article for a volume on rational decision-making. The best thing would be for you to accept the invitation and then write the article. The volume is very prestigious, so having a paper published there will help you to get tenure and receive invitations to conferences. But you believe that if you accept the invitation, you'll wind up procrastinating and never getting around to writing the article. This would be very bad - the editors would be frustrated and you would often be passed over for future volumes. It would be better to reject the invitation than to accept it but not write.

---

<sup>21</sup>This example is a slightly modified version of a case presented in Jackson and Pargetter (1986).

Ought you accept the invitation or decline? It would seem that you ought to decline. After all, you believe that if you were to accept the invitation, you would wind up with the worst possible outcome - accepting the invitation and not writing, which will result in angry editors and decreased opportunities in the future.

But now there is a puzzle, for it also seems that you ought to accept the invitation and write the article. After all, you correctly believe that you are able to accept the invitation and then write the article. And if you were to accept and write, you would wind up with the best possible outcome - far better than what would happen if you were to accept and not write, or if you were to decline and not write.

We have concluded both that you ought not accept and that you ought to accept and write. This is puzzling. To begin with, standard deontic logic (deriving from the work of von Wright (1951)) has the result that  $Ought(A \wedge B)$  entails  $Ought(A) \wedge Ought(B)$  and that  $Ought(A) \wedge Ought(\neg A)$  entails a contradiction. Therefore, if we accept this description of the case, we must give up standard deontic logic (since  $Ought(Accept \wedge Write) \wedge Ought(\neg Accept)$  would entail a contradiction). This is the conclusion Chisholm (1963) draws from this sort of case. Much effort must then be expended in trying to modify standard deontic logic so as to be compatible with this description of your obligations in this case.<sup>22</sup>

Worse, if given your beliefs, it is the case both that you ought to accept and write and that you ought to decline, then you cannot do everything that you ought to do, since obviously you cannot both accept and write and also decline.<sup>23</sup> This is problematic for three reasons. First, it means that our theory is giving you conflicting guidance. Second, it means that you cannot

---

<sup>22</sup>Chisholm puts his case in terms of conditionals, but I prefer for the sake of simplicity to put express it using conjunctions. See footnote 24, below, for the version of the paradox based on conditionals, along with the dissolution of that version of the paradox based on **Options-as-Decisions**.

<sup>23</sup>Jackson and Pargetter (1986) argue that you can in fact do all that you ought to do in this case. But they are considering what you *objectively* ought to do. In considering what you objectively ought to do, whether you ought to accept the invitation depends not on what you believe you will later do, but on what you will in fact later do. Then, in a case where if you accept the invitation, you in fact won't write the review, it appears both that you ought to accept and write, and that you ought to decline. But in this case, it is still possible for you to fulfill all of your requirements, since if you were to accept and write, it would no longer be the case that you ought to have declined. It is only given the truth of the conditional *if you accept, then you won't write* that you ought to decline. But you are able to affect the truth value of that conditional.

avoid being subject to rational criticism, no matter what you do. Third, it throws a wrench into our use of the subjective *ought* to predict agents' behavior. In this case, we obviously would not want to predict both that you will accept and write and that you will decline.

**Options-as-Decisions** yields an attractive dissolution and diagnosis of this paradox. In arriving at this problem, we noted that the best course of action for you in **Professor Procrastinate** is to accept the invitation and write the paper. Because you are also capable of accepting the invitation and then writing the paper (and believe you are capable of so doing), we inferred that you ought to accept the invitation and write the paper.

But if **Options-as-Decisions** is correct, the mistake in this reasoning was to suppose that because you have the ability to perform some action (and believe you have this ability), that action constitutes an option for you and so is potentially something that you ought to do. But we have already seen that the view that everything you are able to do (or believe you are able to do) is an option for you is untenable.

According to **Options-as-Decisions**, your options are all and only the decisions presently open to you - things like (i) *deciding to accept the invitation but not write the paper*, (ii) *deciding to accept the invitation (leaving open the issue of whether to write the paper)*, and (iii) *deciding not to accept the invitation*. (According to the Bratman's account of abilities to make decisions, you cannot make the decision to accept and then write the paper, since you believe that if you were so to decide, you would not carry out the decision. Therefore, this decision is not an option for you.) You are confronted with just this one, unique set of options. And given your belief that if you make a decision that involves accepting the invitation, you won't write the paper, the option with the highest expected utility is *deciding not to accept the invitation*. So you ought to decide not to accept the invitation, *and that's all!* There is no sense in which you ought to decide to accept the invitation and write the paper (if you are even able to make this decisions). (Note that even if making the decision to accept and then write counted as an option for you, it would have low expected utility and hence not be the decision that you ought to make.)

---

But when we are considering the subjective *ought*, things are different. Whether you ought to decline depends not on the actual truth value of the conditional *if you accept, then you won't write*, but on whether you believe that that conditional is true. And while your actions can affect the truth of that conditional, they cannot affect whether you presently believe that conditional to be true.

Chisholm's Paradox arises for Proposals 1-3 precisely because they allow for different ways of chopping up the space of options available to an agent. In **Professor Procrastinate**, they allow us to characterize your options in a coarse-grained way, as consisting of the two options of (i) accepting and (ii) declining, or in a more fine-grained way, as consisting of the three options of (i) accepting and writing, (ii) accepting and not writing, and (iii) declining (and not writing). And so it could be true both that declining has highest expected utility when we characterize your options in a coarse-grained way, and that accepting and writing has highest expected utility when we characterize your options in a more fine-grained way, yielding the result that you both ought to decline and also ought to accept and write. And in general, when a theory of options allows for multiple ways of characterizing your options, there is a possibility of winding up with conflicting *oughts*, since the option with highest expected utility on one way of chopping up your options might be incompatible with the option with highest expected utility on another way of chopping up your options.

**Options-as-Decisions** resolves Chisholm's Paradox by yielding a unique set of options for each agent. There is one set of options for each agent *S* and time *t*, namely the set of propositions of the form *S (only) decides at t to  $\phi$*  that you are able to bring about. As a result, there is no way to wind up with conflicting *ought* claims (and hence no way to wind up with a case that poses a counterexample to standard deontic logic). There is only one set of options, and so whichever proposition has highest expected utility in that one set will be the one that you ought to bring about, period.<sup>24</sup>

---

<sup>24</sup>Chisholm originally presented the paradox using conditionals. The following statements are supposed to all be true descriptions of the case, but they are jointly incompatible with standard deontic logic: (i) You ought to write the paper; (ii) It ought to be that if you write the paper, you accept the invitation; (iii) If you believe you won't write the paper, you ought not accept the invitation; and (iv) You believe you won't write the paper. From (i) and (ii) it follows, by standard deontic logic, that you ought to accept the invitation, while from (iii) and (iv) it follows, by modus ponens, that you ought not accept the invitation. But on standard deontic logic, it cannot be the case both that you ought to accept and that you ought not accept.

But while these statements all *sound* compelling, **Options-as-Decisions** entails that (i) and (ii) are simply false. (i) is false because writing the paper is not an option for you, and (ii) is false because making this conditional true is not an option for you. Chisholm's Paradox shows that an intuitive description of this case (expressed in the statements (i)-(iv)), including a description of your obligations therein, is incompatible with standard deontic logic, given a standard interpretation of the conditionals in (ii) and (iii). One response is to modify standard deontic logic. Another response is to try to reinterpret the

This is a desirable result. But still, you might worry that my account is too lenient. In **Professor Procrastinate**, it says that you ought to decide not to accept the invitation, as a result of your believing yourself to be weak-willed. Does **Options-as-Decisions** therefore have the problematic implication that you are excused from any criticism if you fail to achieve the best possible outcome by accepting and then writing the article?

No. My account allows that you may be still subject to rational criticism if you decide not to accept the invitation, but this criticism will not stem from your having failed to do what you ought to have done. Such criticism can arise in a variety of ways.

First, you might be in fact weak-willed (so that if you were to accept the invitation, you would procrastinate) and take this fact into account in deciding not to accept the invitation. Then, if weakness of will is a failure of rationality, then you are irrational not for having *done* something you rationally ought not have done, but instead for lacking a certain ability, namely the ability to control your future selves through your present decisions.

Second, even if you are not now weak-willed, you might have been weak-willed in the past, and this past weakness of will may have given you the justified but false belief that you are now weak-willed. You responded appropriately to this justified belief in deciding not to accept the invitation, and so your present self is in no way irrational. But if weakness of will is a failure of rationality, you are still subject to rational criticism, although this criticism is directed at your past self rather than your present self.

Third, you might have lacked good evidence for your belief that if you were to accept, you would procrastinate. If this is so, you were not *practically* irrational in deciding not to accept. Rather, you were *epistemically* irrational, since the belief that recommended deciding not to accept was not responsive to your evidence.

However, if you had good but misleading evidence that you are weak-willed, and this evidence was not the result of any past irrationality on your part, then you are in no way subject to rational criticism if you decide not to accept the invitation. Your past self was perfectly rational, you formed beliefs that were supported by your evidence, and you responded appropriately to

---

conditionals in (ii) and (iii) to avoid incompatibility with standard deontic logic. A third response, which falls out of **Options-as-Decisions**, is to simply deny the description of your obligations in this case. If Chisholm's description of your obligations is incorrect, then the paradox dissolves even without any modification of standard deontic logic or non-standard interpretation of the conditionals.

these rational beliefs by deciding to decline the invitation. So, sometimes you take into account predicted weakness of will on your part and therefore wind up with a sub-optimal outcome (among those that you could have achieved if you had performed certain sequences of actions) without being in any way irrational.

## 2.10 Conclusion

The project of coming up with a theory of what an agent's options are is part of a broader project in the theory of rationality. Formal Bayesian tools have proven incredibly fruitful for theorizing about epistemic and practical rationality. On the epistemic side, Bayesian models have illuminated questions about how agents ought to modify their beliefs (or degrees of belief) in response to various sorts of evidence. And on the practical side, Bayesian expected utility theory has shed light on how agents ought to behave, given their uncertainty about the world.

But these formal models embody serious idealizations and are most at home in somewhat artificial casino cases. Suppose you are at the roulette table deliberating about how to bet. In this situation, all of the information needed to employ expected utility theory is right in front of you. Expected utility theory requires three things as inputs in order to yield a recommendation about how to act: probabilities, utilities, and a set of options. At the roulette table, all of these three elements are straightforwardly determined by the setup of the case. The probabilities are precise and given by the physical symmetries of the roulette wheel. The utilities can be thought of as the monetary gains or losses that would be incurred in different outcomes of the spin. And the options are all of the different bets you might place, given the rules of roulette. Expected utility theory is so easy to apply in the casino precisely because it is so clear what the relevant probabilities, utilities, and options are.

But out in the wilds of everyday life, things are not so neat and tidy. Suppose that instead of standing at the roulette table deliberating about to bet, you are standing at the raging creek deliberating about how to act in light of the fast-flowing current. Instead of precise probabilities determined by the physics of the roulette wheel, you might have only hazy degrees of belief about which outcomes would result from various actions you might take. And instead of precise utilities linked to the possible monetary pay-

offs determined by the casino's rules, you might only have rough preferences between possible outcomes (fording the creek safely is better than heading home, which in turn is much, much better than being swept away into the rapids downstream) without being able to assign precise numbers representing their desirabilities. And finally, instead of the options being specified by the rules of roulette, it is *prima facie* less clear which things you should be evaluating as options.

In order for expected utility theory to be relevant to your situation, each of these three disanalogies between the creek case and the roulette case needs to be addressed. How should the formal framework be employed (or extended) in cases where you lack precise probabilities, precise utilities, and a set of options stipulated by the rules of a game? The first of these problems - employing the formal machinery in the absence of precise probabilities - has received considerable attention since the 1970s, and much progress has been made.<sup>25</sup> The second - employing the machinery in the absence of a precise utility function - has only recently been addressed in any detail.<sup>26</sup> But the third - what should count as your options in a case where this isn't specified by something like the rules of the casino - has hardly been dealt with at all.<sup>27</sup>

My aim in this paper has been to confront head-on the problem of coming up with a theory of an agent's options. This task proved surprisingly tricky. On the one hand, your options must be things that you believe you can do and must supervene on your mental states, in order for the subjective *ought* to be appropriately sensitive to your uncertainty about the world. On the other hand, your options must also be things that you can actually do. In a sense, these desiderata require your options to consist of a kind of action such that your actual abilities to perform actions of that kind always match up with your beliefs about your abilities to perform actions of that kind. I argued that only decisions can plausibly play this role. Hence **Options-as-Decisions**, according to which your options consist of all and only the decisions you are presently able to make. Focusing on an agent's decisions, as opposed to the non-mental actions she might be able to perform, impacts on our first-order normative theorizing in a variety of ways. In this paper, I showed how it directly takes into account the costs and benefits of decisions themselves and

---

<sup>25</sup>See especially Levi (1974), Joyce (2005), White (2009), and Elga (2010) for discussion.

<sup>26</sup>See Hare (2010) for compelling discussion of this issue.

<sup>27</sup>Aside from the aforementioned brief discussions in Jeffrey (1965) and Lewis (1981), this issue is discussed in Jackson and Pargetter (1986), Joyce (1999), Pollock (2002), and Smith (2010).

also leads to a principled and attractive response to Chisholm's Paradox.



# Chapter 3

## Options and Diachronic Tragedy

### 3.1 Introduction

In a tragedy, the protagonist suffers some misfortune. What makes this misfortune tragic is that it is foreseeable well before it occurs. In some tragedies, the misfortune is foreseeable only by the audience. But in others the misfortune is in some sense foreseeable by the protagonist himself. The protagonist can foresee that his own desires will drive him to engineer his ruin but nonetheless fails to depart from this disastrous course.

Certain sorts of preferences, of particular interest to philosophers, are tragic in this second way. They drive you to perform each member of a sequence of actions that you can see will result in a bad outcome, even though there is some alternative sequence of actions that you in some sense could have performed and that would have avoided this bad outcome. In this way, these preferences lead you to act over time in a manner that is to your own acknowledged, predictable disadvantage. I call this phenomenon *Diachronic Tragedy*.

I begin with a series of cases of Diachronic Tragedy. Each has generated extensive debate, although these debates have been conducted largely independently of one another. In many cases, philosophers have attacked the preferences involved as irrational simply on the grounds that they yield Diachronic Tragedy.

After discussing these cases and highlighting their common structure, I

attempt to show that this ubiquitous style of argument - concluding that a certain preference is irrational from the premise that it yields Diachronic Tragedy - fails. This style of argument crucially relies on the assumption that the rational *ought* can be applied not only to particular actions, but also to sequences thereof. But given the account of options developed in Chapter 1, it follows that the rational *ought* does not apply to sequences of actions. Therefore, the crucial assumption needed to infer that preferences which yield Diachronic Tragedy are irrational is false.

I conclude that the mere fact that a certain preference yields Diachronic Tragedy does not constitute its being irrational.

## 3.2 Diachronic Tragedy

A *Tragic Sequence* is a sequence of actions  $S_1$  such that at all times you prefer performing some other possible sequence of actions  $S_2$  over performing  $S_1$ . Many different kinds of preferences have the unfortunate consequence that, given those preferences, you will prefer performing each member of a Tragic Sequence at the time it is available, even though you prefer not to perform the sequence as a whole. Such preferences are *Tragic Preferences*. In Sections 2.1-2.7, I present prominent examples of Tragic Preferences from the literature. I want to emphasize that at this point I remain neutral on the question of whether the preferences in the examples to follow are in fact irrational. (The reader may skip some of these examples without loss, as the later discussion will focus on their common core, rather than on the specific details of each case.)

### 3.2.1 Preference Shifts

Shifts in your preferences can be tragic. Suppose you are the Russian Nobleman imagined by Parfit (1986). You are a 20 year old fervent leftist. But you know that by middle age, you will become an equally fervent rightist. Consider:

**The Russian Nobleman:** You will receive an inheritance of \$100,000 at age 60. Right now, you have the option (call it *Donate Early*) of signing a binding contract which will require \$50,000 to be donated to left-wing political causes. No matter whether you take this option, you will at age 60 have the option (call it

*Donate Late*) of donating \$50,000 to right-wing political causes. (No greater donation is permitted under Tsarist campaign finance laws.) Right now, you most prefer donating \$50,000 to left-wing causes and nothing to right-wing causes. But you also prefer donating nothing to either side over donating \$50,000 to each side, as the effects of those donations would cancel each other out.

Right now, regardless of whether your later self will *Donate Late*, you prefer to *Donate Early*.<sup>1</sup> But at age 60, no matter what you do now, you will prefer to *Donate Late*. But the sequence of actions <*Donate Early*, *Donate Late*> is a Tragic Sequence, since at all times, you disprefer it to the sequence <Not *Donate Early*, Not *Donate Late*>. It is better to save your money than to give it all away in two donations that cancel each other out.

### 3.2.2 Time Bias

Time bias is a special case of predictable preference shifts. You are time biased if you prefer that your pains be in the past and your pleasure in the future, even if this means more pain and less pleasure overall. It was widely thought that time bias was a practically inert sort of preference, since you cannot affect the past. But Dougherty (2011) shows that time bias can make a difference to how you act if you are also risk averse, and that it will do so in a way that leads to tragedy. (You are risk averse if you prefer a gamble with a smaller difference between the best and worst possible outcomes to a gamble with a higher difference between *its* best and worst possible outcomes, even if the expected value of the first gamble is somewhat lower than the expected value of the second.)<sup>2</sup>

Suppose you are both time biased and risk averse. Consider:

**Uncertain Pain:** A coin was flipped to determine which of two surgery regimes you will undergo. If it landed heads, you will

---

<sup>1</sup>If your later self does not *Donate Late*, you would rather give \$50,000 to left-wing causes, since this is your most preferred outcome. And if your later self does *Donate Late*, you would rather cancel the effect of that donation by giving \$50,000 to left-wing causes than let that later right-wing donation go unchecked.

<sup>2</sup>So, for instance, you might prefer a bet on a coin toss which pays \$8.50 if heads and \$10.50 if tails to a bet which pays \$0 if heads and \$20 if tails, even though the expected value of the former bet is \$9.50 while the expected value of the latter bet is \$10.

have 4 hours of painful surgery on Tuesday and 1 hour of painful surgery on Thursday (the Early Course). If it landed tails, you will have no surgery on Tuesday and 3 hours of painful surgery on Thursday (the Late Course). Either way, you will be given amnesia on Wednesday, so that you won't remember whether you had surgery on Tuesday (though you will remember everything else). There is a clock next to your bed, so you always know what day it is.

On Monday and Wednesday, you will be offered the pills Help Early and Help Late, respectively. Each reduces the difference between the highest possible amount of future pain and the lowest possible amount of future pain:

**Help Early:** If you are in the Early Course, then taking Help Early will reduce the time of your Thursday surgery by 29 min. If you are in the Late Course, then taking Help Early will increase the time of your Thursday surgery by 31 min.

**Help Late:** If you are in the Early Course, then taking Help Late will increase the time of your Thursday surgery by 30 min. If you are in the Late Course, then taking Help Late will decrease the time of your Thursday surgery by 30 min.

On Monday, you prefer taking Help Early to refusing it. Why? Because it reduces the difference between the highest and lowest amounts of possible future pain (by reducing the future pain in the Early Course scenario involving the most future pain and increasing the future pain in the Late Course scenario involving the least future pain) at a cost of increasing your expected future pain by only 1 min. This is true whether or not you take Help Late.

On Wednesday, you prefer taking Help Late. Why? Because it reduces the difference between the highest and lowest amounts of possible future pain without changing your expected future pain at all. This is true whether or not you took Help Early.

But taking both Help Early and Help Late just guarantees you 1 more minute of pain on Thursday than if you had refused both pills. Hence, the sequence of actions <Take Help Early, Take Help Late> is a Tragic Sequence, since at all times you prefer refusing both pills over taking both pills. Hence, time bias is an example of a Tragic Preference.

### 3.2.3 Intransitive Preferences

Suppose you have intransitive preferences. You prefer Apple Pie to Blueberry Pie, Blueberry Pie to Cherry Pie, and Cherry Pie to Apple Pie. Consider:

**Money Pump:** You start off with an Apple Pie, a Blueberry Pie, and a Cherry Pie. You will be offered three deals in succession no matter what. Deal 1: receive a Blueberry Pie in exchange for your Cherry Pie and 10 cents. Deal 2: receive an Apple Pie in exchange for a Blueberry Pie and 10 cents. Deal 3: receive a Cherry Pie in exchange for an Apple Pie and 10 cents.<sup>3</sup>

If you act on your preferences at each time, you will be turned into a money pump. You will accept the first deal, giving up 10 cents and a Blueberry Pie in exchange for an Cherry Pie. Why? Because regardless of whether you will go on to accept the second and third deals, you would prefer to move up from Cherry Pie to Blueberry Pie, even at a cost of 10 cents. For perfectly analogous reasons, you will accept the second and third deals as well. But having accepted all three deals, you wind up with the same assortment of pies that you started with despite your outlay of 30 cents.

The sequence <Accept Deal 1, Accept Deal 2, Accept Deal 3> is a Tragic Sequence, since throughout the whole process, it is dispreferred to the sequence of declining all three deals. So intransitive preferences are an example of Tragic Preferences.

### 3.2.4 Imprecise Preferences

Suppose your preferences are imprecise - You have no preference between a scuba trip to Australia ( $A$ ) and a safari trip to Botswana ( $B$ ), but you also do not regard them as equally desirable. For adding \$50 to one of them wouldn't make you then prefer it to the other. You don't prefer  $A+$  to  $B$

---

<sup>3</sup>The original money pump argument is due to Davidson et al. (1955). I have presented an improved version of the standard Money Pump case due to Dougherty (ms). In the standard case, where you start off with Cherry, are then given the opportunity to pay to switch to Blueberry and then again to Apple, and then again back to Cherry. The standard case has the disadvantage of being such that you can avoid ruin by simply refusing the first deal, since the later deal (e.g. paying to switch from Blueberry to Apple) cannot be offered unless you accept the first deal (paying to switch from Cherry to Blueberry). Dougherty's case blocks this escape route, since each deal can be offered no matter whether you accept or decline the deals that were offered before.)

or  $B+$  to  $A$  (even though you prefer  $A+$  to  $A$  and  $B+$  to  $B$ ). Imprecise preferences can lead you to misfortune. Consider:

**Scuba or Safari:** There are two boxes. You see that Box A contains a ticket for the scuba trip  $A$ , while Box B contains a ticket for the safari trip  $B$ . You know in advance that at  $t_1$  you will get to decide whether \$50 is placed in Box A or Box B, and then at  $t_2$  you will get to take one of the boxes.

You have no preference about which box to put the \$50 in at  $t_1$  (since the situation is symmetric). Suppose you put the \$50 in Box A. Then at  $t_2$  your preferences license you to take either Box A or Box B. In particular, they license you to take Box B (since you do not prefer scuba plus \$50 over the safari). But the sequence <put \$50 in Box A, take Box B> is a Tragic Sequence, since at all times you prefer the outcome (safari plus \$50) that would have resulted from putting the \$50 in Box B and taking Box B. (Similarly, *mutatis mutandis*, for putting the \$50 in Box B and taking Box A.)<sup>4</sup>

### 3.2.5 Infinite Decisions

One option  $A$  *dominates* another option  $B$  if and only if option  $A$  yields a better outcome than  $B$  in every state of the world. It is widely accepted that you are rationally permitted, and even required, to take dominant options. But Arntzenius et al. (2004) argue that in some infinite cases, taking dominant options will lead to trouble:

**Satan's Apple:** Satan has cut an apple into infinitely many slices. At each of various times  $t_i$ , you are asked whether you would like to eat slice  $\#i$ .<sup>5</sup> If you eat infinitely many slices, you go to Hell, while if you eat only finitely many slices, you go to

---

<sup>4</sup>This argument parallels the argument made by Elga (2010) in the context of imprecise *degrees of belief*. Your degrees of belief are imprecise if there are two propositions  $A$  and  $B$  such that you do not regard one as more likely than the other, but you also do not regard them as equally likely (so that disjoining one with a proposition with small but positive probability does not make you regard the disjunction as more likely than the remaining proposition). Elga shows that having imprecise degrees of belief can license you to perform each member of a Tragic Sequence.

<sup>5</sup>The  $t_i$  are arranged so that the choosing constitutes a supertask -  $t_0$  is 0 sec from now,  $t_1$  is 30 sec from now,  $t_2$  is 45 sec from now, and so on.

Heaven. Your first priority is to go to Heaven rather than Hell.  
Your second priority is to eat as much apple as possible.

For each slice  $i$ , eating that slice dominates not eating it. For eating  $it$  will not make the difference between eating only finitely many slices and eating infinitely many slices, and so it will not affect whether you go to Heaven or to Hell. But if you take the dominant option for each slice - eating it - then you will wind up eating infinitely many slices and be condemned to an eternity in Hell! The sequence of eating every slice is a Tragic Sequence, since it yields a worse outcome than myriad other sequences of actions (e.g. that of refusing every slice). So the preference for dominant options is a Tragic Preference.<sup>6</sup>

### 3.2.6 Formal Epistemology

Formal epistemologists have argued for other principles of rationality on the grounds that violating them can lead you to prefer each member of a Tragic Sequence. I mention three such cases here, without going into full details for reasons of space.

First, Bayesian epistemologists standardly hold that you ought to conditionalize on your evidence. That is, upon learning an evidence proposition  $E$ , you ought to set your degree of belief in every other proposition  $H$  equal to your prior conditional degree of belief in  $H$  given  $E$ .<sup>7</sup> Lewis (1999a) argues for this principle by showing that if you violate conditionalization, you will prefer each member of a Tragic Sequence - a clever bookie could offer you different bets at different times, such that each bet looks good at the time it is offered, even though the bets you accept together guarantee you a loss.

Second, van Fraassen (1984) argues for the principle of Reflection, which states that you ought to defer to the beliefs of your later selves. You ought

---

<sup>6</sup>It is an interesting feature of Satan's Apple that each possible sequence of actions is worse than some other sequence of actions. Therefore, even if you had the ability to decide all at once which slices of apple to eat, it is unclear what you ought to do, since whatever sequence you choose, there will be some better one that you could also have chosen. Perhaps there is some threshold number of slices such that you are permitted to choose any sequence in which you eat at least that many slices (but not infinitely many). But any such threshold will inevitably be arbitrary. Perhaps in this case, we must abandon the binary *ought/ought not* distinction in favor of more graded evaluations, in which we can only speak of an action's being more rational than another.

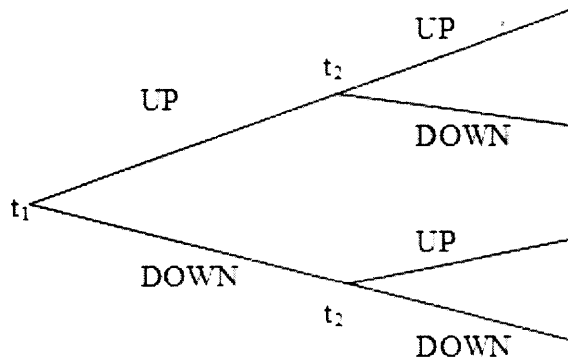
<sup>7</sup>More formally, where  $P_0$  is the probability function representing your initial degrees of belief and  $P_E$  is the probability function representing your degrees of belief after becoming certain of  $E$ , Conditionalization requires that  $P_E(H) = P_0(H|E)$ .

to treat your later selves as experts. You satisfy Reflection if your current degree of belief in  $H$ , conditional on your later having degree of belief  $n$  in  $H$ , is itself  $n$ .<sup>8</sup> van Fraassen argues for Reflection by showing that if you violate it, you will prefer each member of a Tragic Sequence consisting of accepting bets at different times which together guarantee you a loss.

Third, the Sure-Thing Principle in decision theory requires you to prefer  $A$  to  $B$  if and only if you prefer a lottery which has  $A$  as a possible prize to an otherwise identical lottery with  $B$  instead of  $A$  as a possible prize.<sup>9</sup> Raiffa (1968) argues that if you violate the Sure-Thing Principle, you will prefer each member of a Tragic Sequence.

### 3.2.7 Common Structure

In all of these myriad cases, we can represent your decision situation with this tree<sup>10</sup>:



At each of  $t_1$  and  $t_2$ , you can either go UP or DOWN. At the first node, you prefer going UP to going DOWN, no matter what you will later do at  $t_2$ . That

<sup>8</sup>More formally, where  $P_0$  is the probability function representing your degrees of belief at  $t_0$  and  $P_1$  is the probability function representing your degrees of belief at a later time  $t_1$ , Reflection requires that  $P_0(H|P_1(H) = r) = r$ .

<sup>9</sup>More formally, the Sure-Thing Principle requires that for all  $A$ ,  $B$ , and  $C$ :

$$A \geq B \text{ if and only if } \{A, p; C, 1 - p\} \geq \{B, p; C, 1 - p\}$$

where  $\geq$  denotes the *at-least-as-preferred* relation and  $\{A, p; C, 1 - p\}$  is a lottery which yields  $A$  with non-zero probability  $p$  and  $C$  with probability  $1 - p$ .

<sup>10</sup>Of course, the number of decision points required will differ in some cases, like Money Pump (three decision points) and Satan's Apples (infinitely many decision points), but the basic structure is the same.



is, you prefer the sequence <UP, UP> over the sequence <DOWN, UP> and prefer the sequence <UP, DOWN> over the sequence <DOWN, DOWN>. And at  $t_2$ , you prefer going UP to going DOWN, no matter what you did at  $t_1$ . That is, you prefer the sequence <UP, UP> over <UP, DOWN> and prefer the sequence <DOWN, UP> over <DOWN, DOWN>.<sup>11</sup> But at both  $t_1$  and  $t_2$ , you prefer the sequence <DOWN, DOWN> over the sequence <UP, UP>. In this way, the sequence <UP, UP> is a Tragic Sequence, but at each time, you prefer performing the member of this Tragic Sequence available at the time.

Let's go through this with a simple example. In **The Russian Nobleman**, Donate Early plays the role of going UP at  $t_1$  and Donate Late plays the role of going UP at  $t_2$ . Right now, as a young leftist, you prefer to Donate Early, no matter what you will do later at age 60 (that is, you prefer sequences in which you Donate Early to corresponding sequences in which you don't). But at age 60, you will prefer to Donate Late, no matter what you did as a young leftist. But both now and at age 60, you prefer the sequence <Not Donate Early, Not Donate Late> over the sequence <Donate Early, Donate Late>. Similarly, *mutatis mutandis* for all the other cases in Section 2.

In essence, these cases are Prisoner's Dilemmas, in which your  $t_1$  and  $t_2$  time-slices are the two prisoners. In a Prisoner's Dilemma, prisoners A and B must each choose to defect or cooperate. Each prisoner prefers to defect, no matter what the other will do. That is, prisoner A prefers the 'sequence' of actions <A defects, B defects> over <A cooperates, B defects> and prefers <A defects, B cooperates> over <A cooperates, B cooperates>, and similarly, *mutatis mutandis* for prisoner B. But each prisoner prefers <A cooperates, B cooperates> over <A defects, B defects>. In a Prisoner's Dilemma, then, each prisoner prefers performing one action (defecting) no matter what the other will do but would prefer that neither of them perform that action than that both perform it. In cases of Diachronic Tragedy, each time-slice prefers performing some action but would prefer that neither time-slice perform its preferred action than that both perform it.

Crucially, I have been assuming in all these examples that you lack the ability to *self-bind*<sup>12</sup>. That is, there is nothing you can do *now* which will

---

<sup>11</sup>The case of imprecise preferences is slightly different. In that case, you do not actually have the preference at each of  $t_1$  and  $t_2$  for going UP; rather, you just lack the contrary preferences at  $t_1$  and  $t_2$ . I set this detail aside for the sake of clarity.

<sup>12</sup>This term comes from Arntzenius et al. (2004).

causally determine which actions your future time-slices will perform. If you have the ability to self-bind and know it, then if you are rational, you will not perform a Tragic Sequence even if you have Tragic Preferences. This is because, by definition, you prefer performing some other sequence of actions over performing the Tragic Sequence. So, if you know that you can bind yourself to any sequence of actions, you will bind yourself to one of these preferred sequences of actions and thereby avoid performing the Tragic Sequence. In **The Russian Nobleman**, for instance, if in your youth you can bind yourself to a future pattern of donating, you will bind yourself to the sequence of actions <Donate Early, Not Donate Late>, since that is the sequence your youthful self most prefers. (Similarly, if the participants in a Prisoner's Dilemma are able to talk and jointly settle on what they will do, they will each cooperate, since they prefer that they both cooperate rather than both defect.) Because Tragic Preferences only lead to trouble if either lack the ability to self-bind or don't know that you have this ability, I will continue to assume in what follows that you are either unable to self-bind or are ignorant of whether you have this ability.<sup>13</sup>

### 3.3 The No Way Out Argument

As we have seen, many philosophers have concluded from the fact that a certain preference (or set of preferences) yields Diachronic Tragedy that that preference (or set thereof) is irrational. This inference is often taken as intuitively compelling and not in need of further argument. But we should be reluctant to simply trust brute intuition here.

First, we already know that certain attitudes can predictably lead to misfortune without being in any way irrational. For example, having an inflated sense of our own talents keeps us happy, motivated, creative, and successful (Taylor and Brown (1988)). But this does not entail that it is irrational to have accurate, evidence-based beliefs about our own talents! (Of course, the fact that a certain sort of attitude is detrimental may entail

---

<sup>13</sup>One might take the Diachronic Tragedy to show not that Tragic Preferences are irrational, but rather that it is a requirement of rationality that one have the ability to self-bind (and know it); it is a requirement of rationality that one be able to make decisions that bind one's later time-slices to certain courses of action. This line of argument may be supported by the work by Hinchman (2003), Bratman (2007), and Korsgaard (2008), who emphasize the importance to rationality and agency of the capacity to form and execute intentions which guide one's later actions.

that it would be rational to desire not to have that attitude, or to attempt to cause oneself not to have that attitude, but this is quite different from entailing that the attitude itself is irrational.)

Second, there are many, many different ways in which you can be irrational - you might form beliefs irrationally, deliberate irrationally, act irrationally on the basis of rational beliefs and desires, or be irrational in virtue of lacking the ability to self-bind discussed in Section 3. So, even supposing that you are somehow irrational if you perform a Tragic Sequence, further argument is required to show that this irrationality is to be located in your *preferences* rather than in some other fact about you.

Third, it is a curious sociological fact that although in *some* of these cases, philosophers have concluded from Diachronic Tragedy that the relevant preferences are irrational, in others they have been reluctant to draw the same conclusion. Philosophers have argued that Diachronic Tragedy shows that it is irrational to have intransitive preferences, to be time biased, to have imprecise preferences, or to violate Conditionalization, Reflection, or the Sure-Thing Principle. (However, some of these arguments are more highly regarded than others; for instance, Reflection is widely regarded as being subject to numerous counterexamples, which casts doubt on the argument for Reflection based on Diachronic Tragedy.) But they have typically been reluctant to draw the parallel conclusion that it is irrational to change your preferences, to take into account foreseen weakness of will, or to have the preferences at issue in **Satan's Apple**. Of course, this sociological fact in no way entails that an inference from Diachronic Tragedy to irrationality is invalid, but it does mean that we should tread lightly and treat arguments based on Diachronic Tragedy with caution.

Given these concerns, we must look for a compelling *argument* that Tragic Preferences are *ipso facto* irrational.

I suggest that the ground of the intuition that Tragic Preferences are irrational is that having those preferences in some cases means that you cannot help but do something you rationally ought not do.<sup>14</sup> Either you will perform a particular action that you rationally ought not perform, or you will perform a sequence of actions that you rationally ought not perform. You are caught with no way out. This suggests the following argument that

---

<sup>14</sup>Again, the case of imprecise preferences is slightly different, in that the relevant preferences seem to merely *permit*, rather than force, you to do something you rationally ought not do.

such preferences are irrational:<sup>15</sup>

### **The No Way Out Argument**

P1: A set of preferences is irrational if there are cases where no matter what you do, you will have done something that, given those preferences, you rationally ought not have done.

P2: If you have Tragic Preferences, then in some cases no matter what you do you will have done something that, given those preferences, you rationally ought not have done.

C: Tragic Preferences are irrational.

Why believe P1? Rational agents do not do things that they rationally ought not do. So, if having certain preferences entails that no matter what you do, you will have done something you rationally ought not have done, then you cannot be a rational agent and have those preferences. A set of preferences is irrational if you cannot be a rational agent and have those preferences. Hence, a set of preferences is irrational if no matter what you do, you will have done something that, given those preferences, you rationally ought not have done.

Why believe P2? Well, one might argue for P2 by saying that no matter what you do, either you will have performed some particular action you ought not have performed, or you will have performed some *sequence* of actions that you ought not have performed. Consider **The Russian Nobleman**. You rationally ought not perform the sequence of actions <Donate Early, Donate Late>, since at all times you preferred performing some other sequence of actions that was available to you. But you rationally ought to perform the particular action Donate Early, since at the time it is available you prefer to perform it. And you rationally ought to perform the particular action Donate Late, since at the time *it* is available you prefer to perform it. So, you rationally ought to Donate Early, you rationally ought to Donate Late, but you rationally ought *not* perform the sequence <Donate Early, Donate Late>. So you cannot do all that you ought to do. For it is logically impossible for you to Donate Early, Donate Late, but not perform the sequence <Donate Early, Donate Late>. (Similarly for the other cases in Section 2.)

---

<sup>15</sup>While I will later argue on principled, theoretical grounds that this argument is unsound, an initial cause for skepticism is the fact that it does not distinguish among the cases of Tragic Preferences which do seem irrational (like intransitive preferences) and those which do not (like the preferences in **Satan's Apple**).

The crucial assumption in support of P2, then, is that the rational *ought* can be applied not just to particular actions like Donate Early and Donate Late, but also to *sequences* of actions like <Donate Early, Donate Late>. But is this assumption correct? I argue that evaluating this assumption, and therefore P2 itself, requires us to get clear about what your options are in a decision situation. We need a general theory of the sorts of things to which the rational *ought* applies. (I use ‘option’ as a technical term to denote whatever sorts of things are such that you ought or ought not do them in a decision situation, so that it is an analytic truth that if you ought or ought not  $\phi$ , then  $\phi$ -ing is an option for you.)

In the next section, I argue that your options at a time  $t$  consist of all and only the decisions you are able to make at  $t$ . If this is correct, then sequences of actions are not the sorts of things to which the rational *ought* applies. Hence we cannot say that you rationally ought not perform a Tragic Sequence, since this would involve a category mistake. Moreover, it is possible to have Tragic Preferences without doing anything that you rationally ought not do. Hence P2 is false, and The No Way Out Argument fails.

I noted in the previous section that cases of Diachronic Tragedy are essentially intrapersonal Prisoner’s Dilemmas. Essentially, I will be arguing that we should say about these *intrapersonal* Prisoner’s Dilemmas the same thing that is commonly said about standard *interpersonal* Prisoner’s Dilemmas. In the interpersonal Prisoner’s Dilemma, each prisoner prefers to defect. But by each defecting, they wind up with an outcome which by each of their lights is worse than the outcome which would have been achieved had they each cooperated. But this does not mean that they (or their mereological sum) were in any way irrational. I hold that we should say the same thing about the different time-slices of an agent in an intrapersonal Prisoner’s Dilemma. Importantly, I argue not only that we *can* treat the different time-slices of an agent in just the way that we treat the different agents in an interpersonal Prisoner’s Dilemma, but that we are *forced* to do so by an independently motivated account of an agent’s options. To this account of options we now turn.

### 3.4 Options as Decisions

The No Way Out Argument crucially relies on the assumption that the rational *ought* can be applied to sequences of actions. But are sequences of

actions the sorts of things which you rationally ought or ought not do? I use the term ‘options’ to apply to the things to which the rational *ought* can be properly applied, so that the proposition that you ought or ought not  $\phi$  entails that  $\phi$  is an option for you. So, evaluating the assumption that the rational *ought* can be applied to sequences of actions requires determining whether sequences of actions are among your options.

In the previous chapter, I argued for an account of options according to which your options at a particular time consist of all and only the decisions you are able to make at that time. I call this view **Options-as-Decisions**. In this section, I very briefly review the motivations for this view. In the next section, I go on to show how adopting **Options-as-Decisions** leads to a rejection of the No Way Out Argument.

I hold that our central notion of rationality is a *subjective* one, on which the requirements of rationality are sensitive to your perspective on the world. What you rationally ought to do in this sense depends on how you believe the world to be, rather than how it actually is, and on what your preferences are, rather than what is objectively valuable. Suppose that you are thirsty and believe the liquid in the glass to be gin, when it is in fact petrol, as in Williams’ famous example (Williams (1982)). While there may be a sense in which you ought not reach for the glass, I am concerned with the sense of *ought* in which you ought to reach for the glass. This, in my view, is the sense of *ought* that is central to our conception of rationality.<sup>16</sup>

Given this subjective notion of rationality, the rational *ought* is supposed to play three theoretical roles. First, it is to be action-guiding, in the sense that you are at least typically in a position to know what you rationally ought to do. After all, if what you ought to do were something quite inaccessible to you, then our theory of the rational *ought* would be of little help to you in practical deliberation. Second, it is to play an evaluative role, in determining whether you and your actions are criticizable or not on rational grounds.

---

<sup>16</sup>Some philosophers, such as Thomson (1986), deny that there is this subjective sense of *ought* and hold instead that there is only one sense of *ought* - the objective *ought*. In this sense of *ought*, you ought not reach for the glass of liquid in Williams’ case, since the liquid is in fact petrol. Nevertheless, Thomson concedes that it would be ‘sensible’ for you to reach for the glass, given that you believe that it contains gin. She just denies that there is a sense of *ought* that lines up with what it would be ‘sensible’ for you to do. If you side with Thomson in denying that there is a subjective sense of *ought*, you can simply substitute ‘it would be sensible for you to  $\phi$ ’ for each occurrence of ‘you rationally ought to  $\phi$ ’ in what follows.

Third, it is to play a predictive role; against the background assumption that you are rational, we predict that you will do the thing that you rationally ought to do.

I argue that on these theoretical roles to be played by the rational *ought* impose three desiderata on a theory of options, and that these desiderata force us to think of your options as consisting of all and only the decisions you are presently able to make.

**Desideratum 1:** If  $\phi$  is an option for you, you must be able to  $\phi$ .

Why is this? Desideratum 1 is required to avoid violating the principle that *ought* implies *can*. This principle is important for three reasons. First, the rational *ought* gives you poor guidance if it tells you to do something that you cannot do. Second, you are not subject to any rational criticism for failing to do something which in fact you couldn't have done. Third, *ought* implies *can* is essential for the predictive role of the rational *ought*, since we would not want to predict that you would do something which in fact you cannot do.

**Desideratum 2:** If  $\phi$  is an option for you, you must *believe* that you are able to  $\phi$  and that  $\phi$ -ing is completely under your control.

Why is this? If you are in fact able to  $\phi$  but believe that you cannot do so, it is implausible to say that nonetheless you rationally ought to  $\phi$ . Suppose you are hiking along and come to a raging creek. You are in fact able to ford the creek (where fording entails that you make it across), and among the things you are actually able to do, fording the creek is best. But you have serious doubts about whether you are able to ford the creek, since it is swollen with the spring snowmelt and flowing fast. Ought you ford the creek? No. First, the rational *ought* is supposed to be action-guiding, in the sense that you are at least generally in a position to know what you rationally ought to do. But since you are not in a position to know that you are able to ford the creek, you are not in a position to know that fording the creek is among your options (given Desideratum 1), and hence that it is potentially the thing that you rationally ought to do. Second, it would be bizarre to call you irrational if you looked for a different crossing, given your doubts about your ability to ford the creek. Third, we would not want to predict that you would ford the creek, given your beliefs and desires.

Another way to see the need for Desideratum 2 is this: Standard frameworks for ranking options (e.g. expected utility theory) have the feature

that their evaluation of fording, say, cannot take into account the potential bad consequences of trying to ford but failing to do so. This is because the expected utility of fording is the sum of the utilities of possible outcomes, weighted by the probability that that outcome would obtain, conditional on your fording the creek. Therefore, the utility of an outcome in which try but fail to ford the creek makes no difference to the expected utility of fording, since the probability of an outcome in which you try to ford but fail, conditional on your fording the creek, is zero! But your beliefs about what would happen if you were to try to do something but fail in this attempt are often very important for your own deliberations and for our evaluation of you. Desideratum 2 is meant to mitigate the impact of this feature of expected utility theory (that its evaluation of an action does not take into account consequences of trying but failing to perform that action) by restricting your options to things that you believe you are able to do.<sup>17</sup>

**Desideratum 3:** What your options supervenes on your present beliefs and desires.

Why is this? Without Desideratum 3, what you rationally ought to do would likewise fail to supervene on your beliefs and desires. This would be bad for two reasons. First, it would make the rational *ought* insufficiently action-guiding. Because what you rationally ought to do would vary independently of your beliefs and desires, you would not always be in a position to know what you rationally ought to do.<sup>18</sup> Second, it is implausible to evaluate two agents differently (criticizing the one but praising the other) if they

---

<sup>17</sup>Of course, you might believe that you are able to  $\phi$  while still believing that there is some chance that you might try to  $\phi$  but fail. Therefore, just requiring that options be things that you believe you are able to do would not solve the problem at hand. This is why, in the statement of Desideratum 2, I added that you must believe that  $\phi$ -ing is completely under your control in order to  $\phi$  to count as an option. If you believe that  $\phi$ -ing is completely under your control, you don't believe that there is any chance that you might try to  $\phi$  but fail despite your best efforts.

<sup>18</sup>Of course, supervenience of what an agent ought to do on her beliefs and desires is not by itself sufficient for her to be in a position to know what she ought to do. She must also know her beliefs and desires. The important point is that self-knowledge and supervenience of *oughts* on beliefs and desires are individually necessary and jointly sufficient for the agent to be in a position to know what she ought to do. Therefore, if supervenience fails, then even a self-knowing agent would not be in a position to know what she ought to do. Importantly, knowledge of one's beliefs and desires is already required for one to be in a position to know what one ought to do, since even knowing what one's options are, one needs to know what one believes and desires in order to know how to rank those



perform the same action after starting out with the same beliefs and desires, but this is what would be required if we say that *oughts* fail to supervene on beliefs and desires, so that two agents with the same beliefs and desires can differ in which sort of action they ought to perform.<sup>19</sup> Third, the use of the rational *ought* in the prediction of action would yield wildly implausible predictions. Suppose you and your doppelgänger, with the same beliefs and desires, are each facing a raging creek. You are able to ford the creek, but your doppelgänger is not. If options failed to supervene on your beliefs and desires (e.g. by consisting of things that you are physically able to do), we could get the result that while you ought to ford the creek, your doppelgänger ought to do something quite different, like give up and head home. But, given that we predict that rational agents will do what they rationally ought to do, we would then predict that you will ford the creek, while your doppelgänger will do an about-face and head home without so much as getting his feet wet. But this would be bizarre! The two of you are in exactly the same mental state, after all! What could be the source of your radically different behavior? If you immediately performed such very different actions despite being in the very same mental state, it would appear that at least one of you wasn't fully in control of your actions.<sup>20</sup> So Desideratum 3 is needed for the rational *ought* to play its action-guiding, evaluative, and predictive roles.

In the previous chapter, I argued that these three desiderata can only be jointly satisfied by a theory of options on which your options consist only of mental volitional acts of making certain decisions (or, perhaps, forming certain intentions):

---

options. Provided that an agent's options supervene on her beliefs and desires, there are no obstacles to her being in a position to know what her options are that are not already obstacles to her being in a position to know how those options are to be ranked.

<sup>19</sup>This is just to express sympathy with internalism about practical and epistemic rationality. I support internalism about rationality, both practical and epistemic, largely for the standard reason that it seems inappropriate to criticize an agent for holding some belief or performing some action when the agent was in no position to see that she oughtn't hold that belief or perform that action. Moreover, many motivations for adopting externalism about epistemic rationality (respecting the link between justification and truth and responding to skepticism, for example) do not give rise to parallel motivations for adopting externalism about practical rationality.

<sup>20</sup>Note that denying that you or your doppelgänger ought to ford the creek is not to say that either of you ought not ford the creek; *not-ought* does not entail *ought not*. If  $\phi$ -ing isn't an option for you, then it will be the case neither that you ought to  $\phi$  nor that you ought not  $\phi$  (although on my view, it might be that you ought to *decide* to  $\phi$ ).

**Options-as-Decisions:** Your options at  $t$  consist of all and only the actions of the form *You decide at  $t$  to  $\phi$*  that you are able at  $t$  to perform.

If **Options-as-Decisions** is correct, then the rational *ought* applies neither to ordinary non-mental actions like *going to the grocery store* nor to sequences thereof. Instead, it only applies to particular decisions made by particular time-slices of the agent under consideration.

As we shall now see, **Options-as-Decisions** entails the falsity of P2 of The No Way Out Argument.

### 3.5 Decisions and Diachronic Tragedy

According to **Options-as-Decisions**, your options (the things to which the rational *ought* applies) at a time consist of all and only the decisions you are able to make at that time. (These decisions can have contents of very different types; they can be decisions to perform particular actions, decisions to perform sequences of actions,<sup>21</sup> decisions to defer deliberation until a later time, etc.) This means that evaluating whether you acted rationally in performing some sequence of actions requires us to think carefully about what was going on in your head and about the individual decisions you made on the way to performing that sequence of actions. Evaluating whether you were rational over a period of time requires determining whether you were rational in making the decisions you made at each particular time.

Recall P2 of The No Way Out Argument:

P2: If you have Tragic Preferences, then in some cases no matter what you do you will have done something that, given those preferences, you rationally ought not have done.

How does P2 fare in light of **Options-as-Decisions**? Supposing you perform a Tragic Sequence, have you done anything that you rationally ought

---

<sup>21</sup>Indeed, there may not be a principled distinction between particular actions and sequences thereof. The vast majority of the non-mental actions we perform, such as buying groceries, can be thought of either as actions in their own right or as sequences of many smaller actions, such as stepping out the door, walking to the store, putting fruit in the basket, etc. A further attractive feature of **Options-as-Decisions** is that it does not require arbitrary stipulations about the fineness of grain of the actions which should be evaluated in deliberation. All possible decisions count as options, no matter the fineness of grain of their contents.

not have done? The first thing to note is that given **Options-as-Decisions**, we cannot immediately infer that you have done something you rationally ought not have done by performing a Tragic Sequence, since sequences of actions are not among your options (although, importantly, *decisions* to perform certain sequences of actions will be among your options; it's just that the sequences themselves won't count as options). Instead, we must consider how your performing a Tragic Sequence came about, and in particular whether it was the result of any *decision* that you rationally ought not have made. Which decision you ought to make at a particular time depends on whether you believe you will carry it out, or more generally on which outcome you believe will result from your making that particular decision. In this way, the cases discussed in Section 2 were underdescribed, since they did not specify what you believed about how likely you would be to carry out each of the decisions available to you.

On some ways of fleshing out the cases, your performing a Tragic Sequence *might* be the result of a decision that you rationally ought not have made. Return again to **The Russian Nobleman**. Suppose that at the outset, in your youth, you believed that you would carry out whichever decision you made regarding future donations (i.e. you believed you have the ability to self-bind), and moreover this belief was true, but you just decided to Donate Early, deferring until middle-age the question of whether to then Donate Late. Having taken Donate Early, your 60-year-old self then deliberated about that question, decided to Donate Late, and carried out this decision, thus resulting in your having performed the Tragic Sequence <Donate Early, Donate Late>. In this case, your initial decision was irrational. Given your youthful belief that you would carry out whichever decision you then made, the option of just deciding to Donate Early had sub-maximal expected utility. Instead, the option of deciding to perform the sequence <Donate Early, Not Donate Late> had maximal expected utility, since you believed that this decision would result in your Donating Early but not Donating Late, and this was the sequence of actions you most preferred. Therefore, you ought to have decided to <Donate Early, Not Donate Late>. In just deciding to Donate Early, you failed to take advantage of your ability to self-bind, and this was irrational. So it is certainly possible that your performing a Tragic Sequence really did involve your having done something you rationally ought not have done.

But on other ways of fleshing out the cases, you can perform a Tragic Sequence without ever doing anything that you rationally ought not have

done. Here is one such way: You did fully believe that you would carry out whichever decision you made (i.e. you believed yourself to be able to self-bind), but you were in fact wrong about this. Suppose that in your youth, you made the decision to perform the sequence <Donate Early, Not Donate Late>. Given your belief that you would do whatever you decided to do, this was the decision you rationally ought to have made (since you preferred this sequence of actions over any other). But despite having made this decision in your youth and carried out the first part of it (Donating Now), you reopened deliberation later on in life and revised the decision you made in your youth, deciding instead to take Donate Late. This new decision was also one that (having reopened deliberation) you rationally ought to have made, since your older self preferred Donating Later to Not Donating Later. Having carried out this new decision and Donated Later, you wound up performing the Tragic Sequence <Donate Early, Donate Late>. But at no point in the process did you take an option (that is, make a decision) that you rationally ought not have taken, given your beliefs and preferences. The decision you made in your youth to perform the sequence <Donate Early, Not Donate Late> was rational, as was the decision you made at age 60 (having reopened deliberation) to Donate Late. Your performing the Tragic Sequence was the result not of having made any decision that you rationally ought not have made, but of having had a false belief that you would carry out whichever decision you made.

Here is a second way: Your performing a Tragic Sequence was the result of failing to believe that you would carry whichever decision you made (i.e. failing to believe yourself able to self-bind). Suppose that in your youth, you believed that your present decision would make no difference to which action you perform at age 60; it would only determine whether you would Donate Early or not. In this case, three decisions were tied for best: (i) the decision to perform the sequence <Donate Early, Donate Late>, (ii) the decision to perform the sequence <Donate Early, Not Donate Late>, and (iii) the decision to Donate Early (deferring until age 60 the question of whether to then Donate Late). Suppose you make the third decision and carry it out, Donating Early. This decision was perfectly rational in light of your beliefs and preferences. Then, at age 60 you must deliberate anew about whether to Donate Late. Suppose you decide at that point to Donate Late and carry out this decision. This new decision was also perfectly rational in light of your beliefs and preferences at age 60, since you preferred going Donating Late to not doing so. You wind up performing the Tragic Sequence

<Donate Early, Donate Late>, but at no point did you make a decision that you rationally ought not have made. Your decision in your youth to Donate Early was perfectly rational (given your belief that you could not control your 60 year old self), as was your decision at age 60 to Donate Late. Your performing the Tragic Sequence was the result not of having made a decision that you rationally ought not have made, but of having failed to believe that you would carry out whichever decision you made.

In sum, given **Options-as-Decisions**, P2 is false. To begin with, we cannot immediately infer from your performing a Tragic Sequence that you did something you rationally ought not have done, since sequences of actions are not options; they are not the sorts of things to which the rational *ought* applies. Whether your performing a Tragic Sequence involved your doing anything you rationally ought not have done depends on whether the individual decisions leading to that Tragic Sequence were themselves irrational. But in a wide range of cases, in particular those in which you either falsely believed that you would carry out whichever decision you made or else lacked this belief in the first place, your performing a Tragic Sequence can instead be the result of a sequence of perfectly rational decisions. So P2 is false. It is possible to have Tragic Preferences, and even to perform Tragic Sequences, without doing anything that you rationally ought not have done. Therefore, the best argument that Tragic Preferences are irrational, The No Way Out Argument, fails.

According to **Options-as-Decisions**, the rational *ought* applies to particular decisions that can be made by particular time-slices of an agent. Mereological sums of time-slices and sequences of actions are not the proper subjects and objects, respectively, of the rational *ought*. As a result, we cannot infer from Diachronic Tragedy that the preferences of the particular time-slices involved are irrational. As noted earlier, cases of Diachronic Tragedy are essentially *intrapersonal* Prisoner's Dilemmas, and I am in effect arguing that **Options-as-Decisions** forces us to say about cases of Diachronic Tragedy the same thing that is standardly said (e.g. by Parfit (1986)) about the *interpersonal* Prisoner's Dilemma. In the latter, each prisoner prefers to defect, but by each defecting, they wind up with an outcome which by each of their lights is worse than the outcome which would have been achieved had they each cooperated. But this does not mean that they performed an irrational 'group action' or that their mereological sum was irrational, since group actions and mereological sums of people are not proper relata of the rational *ought*. And it does not mean that either individual's preference for

defecting over cooperating was irrational. While treating the time-slices involved in cases of Diachronic Tragedy in the same way that we already treat the people involved in the Prisoner's Dilemma may seem radical, I am arguing that we are forced to do so by an independently motivated account of options.

### 3.6 Is Everything then Permitted?

If **Options-as-Decisions** is a proper characterization of your options, then we are left with no compelling argument that Tragic Preferences are irrational. I draw the stronger conclusion that the fact that certain preferences yield Diachronic Tragedy in no way entails that those preferences are irrational.

Are we therefore committed to thinking that all of the preferences discussed in Section 2 are perfectly rational? Not at all. In some cases, the preferences in question are in fact irrational, albeit for reasons entirely independent of Diachronic Tragedy.

One might, for instance, argue for the irrationality of intransitive preferences along the following lines, loosely inspired by Kolodny<sup>22</sup>. A preference for one thing over another is supported by reasons only if there is more reason to desire the one than to desire the other. But it is a metaphysical truth that 'more,' and hence 'more reason than,' is transitive, and so there cannot be more reason to desire A than to desire B, more reason to desire B than to desire C, and more reason to desire C than to desire A. Therefore, if you have intransitive preferences (e.g. by preferring A to B, B to C, and C to A), the structure of your preferences itself entails that not all of your preferences can be supported by reasons. But if one can read off from the mere structure of your preferences that they cannot all be supported by reasons, then those preferences are irrational. Hence it is irrational to have intransitive preferences.<sup>23</sup>

While this sketch of an argument will of course be controversial, the important thing to note is that there are promising ways of arguing that certain preferences traditionally targetted by appeal to Diachronic Tragedy

---

<sup>22</sup>See especially Kolodny (2007) and Kolodny (2008).

<sup>23</sup>A slightly different argument would appeal to a link between rational preferences and betterness, along with the metaphysical or axiological claim that betterness is transitive.

are in fact irrational, albeit for reasons entirely independent of Diachronic Tragedy.

But with other cases of Tragic Preferences, it is doubtful whether there will be any such independent argument that they are irrational, and we should instead conclude that the preferences are rational despite giving rise to Diachronic Tragedy. This is likely the case, for instance, with the preferences at issue in **Satan's Apple**. All of your preferences are perfectly defensible and coherent, but infinities are strange beasts, and even perfectly rational preferences can land you in trouble if you face infinitely many decision points. As for the other cases, like preference shifts, time-bias, foreseen weakness of will, and imprecise preferences, whether we ought to conclude that they are irrational will likewise depend on whether there is some *independent* argument for their irrationality, perhaps along the lines of the Kolodny-inspired argument against intransitive preferences.

This fact could potentially explain the curious sociological fact mentioned in Section 3, namely that philosophers have been willing to conclude from *some* instances of Diachronic Tragedy that the relevant preferences are irrational but have been reluctant to draw the same conclusion in other instances of Diachronic Tragedy. It may be that in the instances where philosophers have been willing to use Diachronic Tragedy to argue for the irrationality of the preferences involved, there is an independent Kolodny-style argument that those preferences are irrational, whereas in the instances where philosophers have been reluctant to draw the same conclusion, this is because there is no such independent argument for the irrationality of the relevant preferences. Perhaps intuitions about rationality or irrationality in cases of Diachronic Tragedy are really tracking whether there is some such independent argument that could support the charge that the preferences involved are irrational.

But whether or not this psychological speculation about the source of our intuitions is correct, it remains the case that we *ought* to evaluate the rationality of Tragic Preferences on independent grounds; just pointing out that they yield Diachronic Tragedy is not enough.

### 3.7 Conclusion

There are myriad cases where having certain preferences predictably leads you to perform sequences of actions that are to your own acknowledged dis-

advantage. They yield Diachronic Tragedy. But the preferences themselves have little else in common. Some philosophers have placed great weight on this single shared feature and concluded that certain Tragic Preferences are irrational. I have argued that on a proper conception of what your options are, the best *argument* for this conclusion fails, and that we have no reason to think that Tragic Preferences are *ipso facto* irrational.

This single shared feature of the preferences discussed in Section 2 should not blind us to their many differences. In some cases (e.g. intransitive preferences), the preferences really are irrational, but this irrationality is identifiable independently of any considerations about sequences of actions or predictable exploitability. But in many other cases, the preferences are perfectly rational. They are defensible and supported by reasons at all times despite their potential to lead to misfortune.

Diachronic Tragedy is a symptom. But like many symptoms, it can result from a variety of different diseases, or even in the absence of any disease at all. In some cases, Diachronic Tragedy is the symptom of irrational preferences, but in others, Diachronic Tragedy arises in the absence of any irrationality at all. Therefore, to give a proper diagnosis of a case of Diachronic Tragedy, we must look beyond this superficial symptom and direct our attention to the underlying preferences themselves. We must determine on independent grounds whether the preferences are defensible and supported by reasons, rather than merely noting that they yield Diachronic Tragedy.



# Chapter 4

## Incoherence Without Exploitability

### 4.1 Introduction

If you believe that your sister is in Tashkent and that Tashkent is in Uzbekistan, but you also believe that your sister is not in Uzbekistan, then your beliefs are not merely odd, but irrational. Moreover, it seems your beliefs are irrational in virtue of their having a certain structure, in particular their being logically inconsistent. It is a requirement of rationality that your beliefs be logically consistent.

But belief is not an all or nothing affair. It comes in degrees. We believe some things more strongly than others. Your levels of confidence, or degrees of belief, are known as *credences*. Are there any rational constraints on the structure of your credences, just as there are rational constraints on the structure of your binary beliefs? Suppose you are certain that your sister is in Tashkent, and you regard it as rather likely that Tashkent is in Uzbekistan, but you also regard it as unlikely that your sister is in Uzbekistan. Your credences are in some sense incoherent and resemble inconsistent beliefs.

Plausibly, you are irrational if you have these incoherent credences, just as you would be irrational in virtue of having inconsistent beliefs. But can we give any *argument* that this sort of incoherence is irrational?

The Dutch Book Argument (DBA) is the most prominent argument that purports to show that it is indeed a requirement of rationality that your

credences have a certain structure.<sup>1</sup> It is a requirement of rationality that your credences be *coherent*, in a sense to be made precise in the next section. For if your credences are incoherent, then you will be predictably exploitable by a clever bookie who knows no more about the world than you do. A Dutch Book is a set of bets that together guarantee you a loss. The idea, then, is that if your credences are incoherent, they will license you to accept each member of some Dutch Book. In this way, a bookie could exploit you by getting you to accept each bet in a Dutch Book, thereby guaranteeing himself a gain and you a loss.

There are many objections one might raise to the DBA. Is it irrational to be predictably exploitable? And even if it is irrational to be predictably exploitable, is this irrationality distinctively *epistemic*, rather than merely pragmatic? What if you have an aversion to gambling or don't care about money, so that you wouldn't want to engage in any betting behavior whatsoever?

Defenders of the DBA have presented thoughtful responses to all of these objections.<sup>2</sup> But there is a deeper problem with the DBA. I will argue that, even if you have no aversion to gambling and care only about money, it is possible to have incoherent credences without being predictably exploitable, provided that your credences are incoherent in a very particular way. Therefore, even if we accept that predictable exploitability is irrational, the DBA can at best show that *some* incoherent credences are irrational. It cannot be used to condemn all incoherent credences as irrational. As such, it is only a partial result. I close by arguing that if it is possible to be incoherent at all, then some structural constraints on credences must be brute requirements of rationality that we accept on the basis of their intuitive plausibility rather than on the basis of any compelling argument.

---

<sup>1</sup>Actually, there are a variety of Dutch Book arguments. My focus here is on the Dutch Book Argument for conformity to the probability calculus. My comments will not bear on, for example, the diachronic Dutch Book arguments for conditionalization (Lewis (1999a), reported earlier by Teller (1973)) and reflection (van Fraassen (1984)).

<sup>2</sup>See in particular Christensen (1996), who argues that the thrust of the DBA is not that if you have incoherent credences, you might lose money, but that if you have incoherent credences, you are inconsistent in how you evaluate the fairness of bets, and this is a distinctively epistemic sort of inconsistency. So, it is immaterial whether there are actually any clever bookies around, or whether you actually dislike gambling. Skyrms (1987) suggests a similar interpretation of the argument. My comments on the DBA will apply equally to their non-pragmatic interpretation.

## 4.2 The Canonical Dutch Book Argument

To make the DBA precise, we first have to say more about what we mean by saying that certain credences are coherent or incoherent. Assuming that your credences can be represented by numbers, they are said to be coherent just in case they conform to the probability calculus. That is, given a set  $S$  of propositions (taken to be sets of possible worlds) closed under complementation and union, your credences are coherent if and only if they obey the following three axioms<sup>3</sup>:

**Non-Negativity:**  $P(A) \geq 0$  for all  $A \in S$

**Normalization:**  $P(T) = 1$  for all necessary truths  $T \in S$

**Finite Additivity:**  $P(A \vee B) = P(A) + P(B)$  for all disjoint  $A, B \in S$

Next, we have to be more precise about the notion of predictable exploitability and what it is for your credences to license you to perform certain actions. To say that your credences license an action is just to say that, supposing those credences were rational, it would be rationally permissible for you to perform that action. You are predictably exploitable just in case your credences license you to accept each member of some Dutch Book.

Then, we can put the DBA in the following form:

**P1:** If your credences are incoherent, then they license you to accept each member of a Dutch Book.

**P2:** If your credences license you to accept each member of a Dutch Book, then it is irrational to have those credences.

**C:** It is irrational to have incoherent credences.

Why believe Premise 1? Here, the DBA focuses on the notion of a *fair betting quotient*. Your fair betting quotient for  $A$  is defined as the number  $n$  such that your credences license you to accept either side of a bet on  $A$  at  $n : 1 - n$  odds (i.e. to accept a bet which pays  $\$(1 - n)$  if  $A$  and  $\$(-n)$  if  $\neg A$  and also a bet which pays  $\$(n - 1)$  if  $A$  and  $\$n$  if  $\neg A$ ).<sup>4</sup>

---

<sup>3</sup>This is using the Kolmogorov (1933) axiomatization.

<sup>4</sup>Sometimes, the notion of a fair betting quotient is explicated in more behavioristic terms, as the number  $n$  such that you are in fact disposed to accept either side of a bet on the relevant proposition at  $n : 1 - n$  odds. As behaviorism has fallen out of favor, I explicate the notion of a fair betting quotient in normative terms, as involving the odds at which your credences license you to take either side of the bet.

Then, the DBA supports Premise 1 with (i) an assumption that your fair betting quotients match your credences (so that for all  $A$  and  $n$ , having credence  $n$  in  $A$  licenses you to accept a bet which pays  $\$(1 - n)$  if  $A$  and  $\$(-n)$  if  $\neg A$  and also a bet which pays  $\$(n - 1)$  if  $A$  and  $\$n$  if  $\neg A$ ), and (ii) the Dutch Book Theorem, proven by de Finetti (1937), which states that if your credences give rise to fair betting quotients that are incoherent (i.e. are negative, non-finitely-additive, or not 1 for each necessary truth), then they license you to accept all the bets in some Dutch Book. The Dutch Book Theorem goes as follows: <sup>5</sup>

### **Dutch Book Theorem for Non-Negativity**

Suppose that your fair betting quotient for  $A$  is  $n$ , where  $n < 0$ . Then, by definition, you are licensed to accept a bet which pays you  $\$(n - 1)$  if  $A$  and  $\$(n)$  otherwise. But this bet guarantees a loss of at least  $\$(-n)$ .

### **Dutch Book Theorem for Normalization**

Suppose that your fair betting quotient for some necessary truth  $T$  is  $n$ , where  $n < 1$ . Then, by definition, you are licensed to accept a bet which pays  $\$(n - 1)$  if  $T$  and  $\$n$  otherwise. But since  $T$  is a necessary truth, this bet guarantees you  $\$(n - 1)$ , which is negative.

Now, suppose your fair betting quotient for  $T$  is  $m$ , for some  $m > 1$ . Then, you are licensed to accept a bet which pays  $\$(1 - m)$  if  $T$  and  $\$(-m)$  otherwise. But again, this guarantees you  $\$(1 - m)$ , which is negative.

### **Dutch Book Theorem for Finite Additivity**

Suppose that your fair betting quotient for  $A$  is  $x$ , your fair betting quotient for  $B$  is  $y$ , and your fair betting quotient for  $A \vee B$  is  $z$ , where  $A$  and  $B$  are disjoint. If  $z < x + y$ , then you are licensed to accept a bet which pays  $\$(1 - x)$  if  $A$  and  $\$(-x)$  otherwise, a bet which pays  $\$(1 - y)$  if  $B$  and  $\$(-y)$  otherwise, and a bet which pays  $\$(z - 1)$  if  $A \vee B$  and  $\$z$  otherwise. These bets together yield  $\$(-x - y + z)$  no matter what, but since  $z < x + y$ , this is negative and hence a loss for you.

---

<sup>5</sup>Here my explication closely follows that of Hajek (2008).

Suppose instead that  $z > x + y$ . Then you are licensed to accept a bet which pays  $\$(x - 1)$  if  $A$  and  $\$x$  otherwise, a bet which pays  $\$(y - 1)$  if  $A$  and  $\$y$  otherwise, and a bet which pays  $\$(1 - z)$  if  $A \vee B$  and  $\$(-z)$  otherwise. These bets together yield  $\$(x + y - z)$  no matter what, but since  $z > x + y$ , this is negative and hence a loss for you.

The Dutch Book Theorem and the assumption that your fair betting quotients match your credences together entail Premise 1.

Premise 2 is pretheoretically compelling. There seems to be something irrational about credences which license you to accept a set of bets that logically guarantees you a loss. But there would be a worry about Premise 2 if *all* credences, whether coherent or incoherent, licensed you to accept Dutch Books. However, the Converse Dutch Book Theorem (proven by Kemeny (1955) and Lehman (1955)) states that if your credences are coherent, then they do not license you to accept any Dutch Books.

To summarize, Premise 1 is supported by the assumption that fair betting quotients match credences and the Dutch Book Theorem. Premise 2 is supported by intuition and the Converse Dutch Book Theorem (which forestalls a potential objection to Premise 2). Premises 1 and 2 together entail the conclusion that it is not consistent with your being rational that you have incoherent credences.

### 4.3 How Credences License Actions

The DBA relies on specific claims about the bets that your credences license you to accept. How should we evaluate these claims? We have a natural, compelling general framework for thinking about which actions, including those related to betting, your credences license you to perform. This framework is known as expected utility theory. Given a set of options available to you, expected utility theory says that your credences license you to choose the option with the highest expected utility, defined as:

$$EU(A) = \sum_i P(O_i|A) \times U(O_i)^6$$

---

<sup>6</sup> $P$  is your credence function,  $U$  is your utility function, representing your preferences, and the  $O_i$  are outcomes that form a partition of logical space. This is the formula for Evidential Decision Theory. I ignore Causal Decision Theory here since it is more com-

On this view, we should evaluate which bets your credences license you to accept by looking at the expected utilities of those bets. Ignoring worries about aversions to gambling and indifference to money (so that your utility function is linear with respect to money), your credences license you to accept a bet just in case that bet has non-negative expected value, given your credences. This is because the alternative option, rejecting the bet, has an expected value of 0, so the option of accepting the bet has an expected value at least as great as that of its alternative just in case the bet's expected value is non-negative. Then, to say that your credences license you to accept each bet in a Dutch Book is to say that each bet in the Dutch Book has non-negative expected value, given your credences. And the claim that incoherent credences license you to accept Dutch Books is the claim that, for each incoherent credence function, there is a Dutch Book, each member of which has non-negative expected value, given that credence function.

As a quick example of how this works, consider Normalization. Suppose you violate Normalization by having credence 0.8 in a necessary proposition  $T$  and credence 0.2 in its negation. Then, consider a bet which pays \$ - 0.20 if  $T$  and \$0.80 if  $\neg T$ . The expected value of this bet is  $(0.8 \times \$ - 0.20) + (0.2 \times \$0.80) = \$0$ . Since this bet has non-negative expected value, you are licensed to accept it. But since  $T$  is a necessary truth, the bet guarantees you a loss of \$0.20.

Now, one might object to this use of expected utility considerations by claiming that expected utility theory only applies to agents with probabilistically coherent credences. Talk of *expected* utility presupposes that we are dealing with a coherent probability function, since the mathematical expectation of a random variable (such as utility) by definition requires the use of a coherent probability function. So, if your credences are incoherent, it doesn't even make sense of talk about the expected utility of an action, relative to your credences.

If this were right, then the DBA would be in serious trouble, since it is unclear how credences could license accepting certain bets in a manner independent of expected utility theory. But I think the objection is misguided. It is not as though the expected utility formula itself requires that we plug in only coherent credences in order to get an output. When we plug in negative credences, or credences greater than one, and the like, we still get a

---

plex, and the distinction between Evidentialism and Causalism is immaterial for present purposes.

number; we don't end up dividing by zero or anything like that. So even if it is improper to use the term 'expected utility' in the context of an incoherent credence function, we can still talk about the *schmexpected utility* of an action  $A$ , which is just the result of plugging in credences to the formula  $\sum_i P(O_i|A) \times U(O_i)$ , whether or not they are coherent. This is just expected utility, minus the presupposition that we're dealing with a coherent credence function. And importantly, the motivations for employing expected utility theory in the context of coherent credences carry over to employing schmexpected utility theory in the context of incoherent credences. In general, we should judge actions by taking the sum of the values of each possible outcome of that action, weighted by one's credence that the action will result in that outcome. This is a very intuitive proposal for how to evaluate actions that applies even in the context of incoherent credences.

From now on, therefore, I will be assuming that expected utility theory is the last word on how credences license you to accept bets.

## 4.4 Negation Coherence and the Limits of the DBA

We saw earlier that a crucial assumption of the DBA is that your fair betting quotients match your credences; i.e. that having credence  $n$  in  $A$  licenses you to accept a bet which pays  $\$(1 - n)$  if  $A$  and  $\$(-n)$  if  $\neg A$  and a bet which pays  $\$(n - 1)$  if  $A$  and  $\$n$  if  $\neg A$  (the other side of the former bet). But given expected utility theory, this assumption fails. We should think of your fair betting quotient for  $A$  as the number  $n$  such that a bet which pays  $\$(1 - n)$  if  $A$  and  $\$(-n)$  if  $\neg A$  and a bet which pays  $\$(n - 1)$  if  $A$  and  $\$n$  if  $\neg A$  each has an expected value of 0 (so that you are licensed to accept either). The expected value of the former bet is:

$$P(A)(1 - n) + P(\neg A)(-n)$$

And the expected value of the latter bet is:

$$P(A)(n - 1) + P(\neg A)(n)$$

Each of these expected values equals 0 when  $n = P(A)/(P(A) + P(\neg A))$ . So, your fair betting quotient for  $A$  is  $P(A)/(P(A) + P(\neg A))$ :

$$\text{Fair Betting Quotients: } FBQ(A) = P(A)/(P(A) + P(\neg A))$$

This means that your fair betting quotients match your credences just in case your credences in a proposition and its negation sum to 1 (that is,  $FBQ(A) = P(A)$  just in case  $P(A) + P(\neg A) = 1$ ). Call this constraint ‘Negation Coherence’:

**Negation Coherence:** For all  $A$ ,  $P(A) + P(\neg A) = 1$

Therefore, the assumption that your fair betting quotients match your credences presupposes that you satisfy Negation Coherence. If you violate Negation Coherence, you might have credence  $n$  in some proposition without being licensed to accept each side of a bet on that proposition at  $n : 1 - n$  odds.

Consider the example from the previous section, where you violate Normalization by having credence 0.8 in a necessary truth  $T$ . If you satisfy Negation Coherence and therefore also have credence 0.2 in  $\neg T$ , then a bet which pays \$ - 0.20 if  $T$  and \$0.80 if  $\neg T$  has an expected value of \$0, and so your credences license you to accept this bet, even though it guarantees you a loss. But if you violate Negation Coherence and assign, say, credence 0.1 to  $\neg T$ , then this bet has an expected value of  $(0.8 \times \$ - 0.20) + (0.1 \times \$0.80) = \$ - 0.08$ , and hence your credences do not license you to accept this bet.

The DBA cannot presuppose Negation Coherence without circularity, since Negation Coherence is a consequence of probabilistic coherence. In particular, it follows from Normalization and Finite Additivity.<sup>7</sup> The DBA can still be used to argue that if you satisfy Negation Coherence, then any incoherence elsewhere in your credence function is irrational (i.e. that it is a requirement of rationality that if you satisfy Negation Coherence, then you fully satisfy probabilistic coherence), but it cannot itself show that Negation Coherence is a requirement of rationality.

## 4.5 Incoherence without Exploitability

Not only does the assumption that fair betting quotients match credences fail, but there are also cases where your credences are incoherent and yet there is no set of bets, each of which has a non-negative expected value, which together guarantee you a loss. Hence, you are provably invulnerable

---

<sup>7</sup>By Normalization,  $P(A \vee \neg A) = 1$ . Since  $A$  and  $\neg A$  are disjoint, Finite Additivity requires that  $P(A) + P(\neg A) = P(A \vee \neg A)$ .



to Dutch Books. Taking propositions to be sets of possible worlds, consider the following credences in a model with three worlds,  $w_1, w_2$ , and  $w_3$ .

$P(\{w_1, w_2, w_3\})$	= 0.9	(Your credence in the necessary proposition)
$P(\emptyset)$	= 0	(Your credence in the necessary falsehood)
$P(\{w_1\})$	= 0.7	(Your credence that you are in $w_1$ )
$P(\{w_2, w_3\})$	= 0	(Your credence that you are not in $w_1$ )
$P(\{w_2\})$	= 0	(Your credence that you are in $w_2$ )
$P(\{w_1, w_3\})$	= 0.8	(Your credence that you are not in $w_2$ )
$P(\{w_3\})$	= 0	(Your credence that you are in $w_3$ )
$P(\{w_1, w_2\})$	= 0.8	(Your credence that you are not in $w_3$ )

The first thing to note is that these incoherent credences yield coherent fair betting quotients. To see this, note that whenever  $P(A) \neq 0$  and  $P(\neg A) = 0$ , your fair betting quotient for  $A$  is 1 and your fair betting quotient for  $\neg A$  is 0. This is because, as we saw in the previous section, your fair betting quotient for  $A$  is  $P(A)/(P(A) + P(\neg A))$ , while your fair betting quotient for  $\neg A$  is  $P(\neg A)/(P(\neg A) + P(A))$ . Hence, your fair betting quotients in the case above are:

$FBQ(\{w_1, w_2, w_3\})$	= 1
$FBQ(\emptyset)$	= 0
$FBQ(\{w_1\})$	= 1
$FBQ(\{w_2, w_3\})$	= 0
$FBQ(\{w_2\})$	= 0
$FBQ(\{w_1, w_3\})$	= 1
$FBQ(\{w_3\})$	= 0
$FBQ(\{w_1, w_2\})$	= 1

These fair betting quotients are non-negative, finitely additive, and 1 for the necessary proposition. As a result, the Dutch Book Theorem does not kick in to entail that you are licensed to accept all the bets in a Dutch Book in this case.

We can go further and prove that these credences do not license you to accept any Dutch Books.<sup>8</sup> One way to prove that you are not licensed to

---

<sup>8</sup>Unfortunately, the Converse Dutch Book Theorem does not itself entail that you are not licensed to accept any Dutch Books in this case. This is because the theorem does not state that if your fair betting quotients are coherent, then you are not licensed to accept a Dutch Book. Rather, it states that if your fair betting quotients are coherent *and* match

accept any Dutch Books is to show that there is a world such that you are not licensed to accept any bet which has you losing money if that world is the actual world; that is, that there is a world such that no bet which costs you money in that world has an expected value greater than or equal to 0. In this model,  $w_1$  is such a world. This is because any proposition  $P$  which is true in  $w_1$  is such that you assign positive credence to it and zero credence to its negation. Hence, you are not licensed to accept any bets on  $P$  which cost you money if  $P$  is true, since in the expected value calculation, losses incurred if  $P$  is true could not be compensated for by gains made if  $\neg P$  is true.<sup>9</sup> So, you are not licensed to accept any bet which costs you money if  $w_1$  is the actual world. Hence, you are not licensed to accept a set of bets which costs you money *no matter which world is the actual world*. Hence, you are not licensed to accept a Dutch Book.

As a final point, it was important that in the case above, you did not have any negative credences; you only violated Finite Additivity and Normalization. This is because we can prove in general that if you have negative credences, then you are licensed to accept a Dutch Book. The reason is that negative credences essentially flip losses to gains in the expected value calculations. Suppose that  $P(A) < 0$ , leaving open whether  $P(\neg A)$  is less than, greater than, or equal to 0. Consider a bet which costs you  $\$r$  if  $A$  and costs you  $\$s$  if  $\neg A$ . By making  $r$  sufficiently large (so that  $P(A) \times (-r)$  is large and positive) and making  $s$  sufficiently small (so that the absolute value of  $P(\neg A) \times (-s)$  is less than  $P(A) \times (-r)$ ), we can make this bet have positive expected value for you, despite the fact that it guarantees you a loss. Nevertheless, the case above shows that you can violate both Finite Additivity and Normalization and yet not be licensed to accept any Dutch Books.

---

your credences, then you are not licensed to accept a Dutch Book. In this case, where your fair betting quotients are coherent but don't match your credences, neither the Dutch Book Theorem nor its converse applies. Moreover, it is possible to have incoherent credences that yield coherent fair betting quotients but still lead to exploitability. For instance, in the example above, multiplying all your credences by  $-1$  would yield a credence function involving negative credences but which still yields the same coherent fair betting quotients. However, as I show below, having negative credences guarantees exploitability. So just having coherent fair betting quotients is not sufficient to avoid exploitability.

<sup>9</sup>For instance, you are not licensed to accept a bet on  $\{w_1, w_2, w_3\}$  which costs you money if  $\{w_1, w_2, w_3\}$  is true, since you assign positive credence to that proposition and zero credence to its negation, and so in the expected value calculation, losses incurred if  $\{w_1, w_2, w_3\}$  is true could not be compensated for by gains made if its negation,  $\emptyset$ , is true. Similarly for  $\{w_1\}$  and its negation,  $\{w_2, w_3\}$ , and for  $\{w_1, w_2\}$  and its negation,  $\{w_3\}$ .

## 4.6 An Objection: Credences Constitutively Linked to Fair Betting Quotients?

I have been employing expected utility theory to determine which actions, including actions related to betting, are licensed by certain credences. A credence function licenses you to take a bet, or perform any other sort of action, just in case that bet or action has highest expected utility. This is the case even for incoherent credence functions. Given this decision-theoretic perspective, fair betting quotients needn't match credences when those credences violate Negation Coherence, and some violations of Negation Coherence yield invulnerability to Dutch Books.

One might object to this argument and defend the DBA by rejecting this blanket application of expected utility theory. One might argue that credences are in fact constitutively linked to fair betting quotients, so that it's part of how we should use the term 'credence' that fair betting quotients always match credences.<sup>10</sup> On this sort of view, having credence  $n$  in  $A$  always licenses you to accept either side of a bet on  $A$  at  $n : 1 - n$  odds, independently of any facts about expected utilities.

If you accept this view, then my objections to the DBA do not apply; different views about credences yield different results about betting behavior. But for my part, I prefer a conception of credences and their relation to action on which betting behavior is seen as just one type of behavior among many. I see no motivation for privileging betting behavior in particular. Both the decision-theoretic view I espouse and a view on which there is a constitutive connection between fair betting quotients and credences yield a tight link between credences and action. But a fully decision-theoretic view, on which credences license all types of behavior, whether betting-related or not, in the same way, is much more unified than a view on which credences license certain actions (accepting certain bets) in a manner independent of expected utility theory, and license other actions in a manner determined by expected utility theory. A view on which the relationship between credences and the actions they license is fully captured by a general decision theory is simpler

---

<sup>10</sup>This idea has its roots in de Finetti (1937), albeit with a more behavioristic understanding of fair betting quotients. For de Finetti, your credence in  $A$  is  $n$  just in case you are actually willing (as opposed to merely licensed) to accept either side of a bet on  $A$  at  $n : 1 - n$  odds. The claim that there is a constitutive connection between fair betting quotients and credences is also found in contemporary discussions of the DBA, such as Maher (1993) and Howson and Urbach (1993).

and more attractive than one which credences can license different actions in very different ways.

## 4.7 Where do We Go From Here?

I have shown that a careful look at how credences license actions reveals the limitations of the DBA. The DBA cannot be used to condemn violations of Negation Coherence as irrational. Worse, incoherent credence functions can provably fail to license accepting any Dutch Books, provided that they violate Negation Coherence.<sup>11</sup> Hence, the DBA can at best show that, once you satisfy Negation Coherence, any deviation from probabilistic coherence elsewhere in your credence function is irrational.

Therefore, in order to have an argument that all rational agents have coherent credences, we must at least supplement the DBA with a separate argument that all rational agents satisfy Negation Coherence. How might this be done?

### 4.7.1 Impossible to Violate Negation Coherence?

One approach would be to argue that it is in fact *impossible* to violate Negation Coherence, that it is a prerequisite for being an agent that one satisfy Negation Coherence. Then, one might say that all rational agents have coherent credences, where this is partly due to their *agency*, which ensures that they satisfy Negation Coherence, and partly due to their *rationality*, which ensures that they go on to fully satisfy coherence. Agency itself gets us to Negation Coherence, and the DBA gets us from there to full coherence.

As an analogy, consider Non-Negativity. Arguably, there is not a distinct norm of rationality that one not have negative credences. Rather, it is simply impossible to have such credences. It is a conceptual truth that credences (unlike utilities, say) have upper and lower bounds, since one's degree of confidence cannot exceed absolute certainty, and one's degree of confidence cannot fall below absolute certainty of falsehood. As a partly conventional matter, we choose to represent the lower bound with 0.<sup>12</sup> Therefore, one

---

<sup>11</sup>To be clear, it is not the case that all incoherent credence functions which violate Negation Coherence are invulnerable to Dutch Books. Only some are.

<sup>12</sup>There are, however, good reasons for using 0 as the lower bound on credences. In principle, we could set the lower bound on credences below 0, but this would require

cannot have negative credences simply because one's credence cannot be lower than absolute certainty of falsehood.

Might it also be impossible to violate Negation Coherence? Offhand, it seems possible to violate Negation Coherence, and not just due to thoughtlessness or stupidity. One might, for instance, adhere to a non-classical logic.

In intuitionistic logic, an atomic proposition is true only if there is a proof of that proposition, and the negation of a proposition is true only if there is a *reductio ad absurdum* of that proposition. Thus, the law of excluded middle fails, since there might be neither a proof nor a *reductio* of some given proposition. An adherent of intuitionistic logic might therefore assign some credence  $n < 1$  to  $A \vee \neg A$  and credences in  $A$  and  $\neg A$  that sum to  $n$ . This person therefore violates Negation Coherence. (Assuming that intuitionism is in fact false, this person also violates Normalization, since  $A \vee \neg A$  is in fact a necessary truth.)

In paraconsistent logic, famously defended by Priest (1979), it is possible for a proposition and its negation to both be true. So,  $A \wedge \neg A$  is not a contradiction. But the law of excluded middle still holds. An adherent of paraconsistent logic might therefore assign some some credence  $n > 0.5$  to  $A$  and credence  $m > 0.5$  to  $\neg A$ , while still assigning credence 1 to  $A \vee \neg A$ . This person therefore violates Negation Coherence as well. (This person also violates Finite Additivity, since  $P(A \vee \neg A) \neq P(A) + P(\neg A)$ .)<sup>13</sup>

If this is right, then it is possible to violate Negation Coherence. And so the fact that the DBA cannot show that Negation Coherence is a requirement of rationality really is an important limit on the role the DBA can play in epistemology. In particular, it cannot rule out the acceptance of a non-classical logic as irrational.

Of course, this does not completely settle the matter, since not everyone will agree with the intuitive take on the credences of intuitionists and paraconsistent logicians I gave above. One might argue, for instance, that the intuitionist does not really doubt the truth of  $A \vee \neg A$ , but instead just doubts the truth of the contingent, metalinguistic proposition that the sentence ' $A \vee \neg A$ ' expresses a truth (and similarly, *mutatis mutandis*, for the paraconsistent logician). Such a philosopher might then go on to hold that it is in fact impossible to violate Negation Coherence, perhaps by arguing that

---

considerable complexity elsewhere in the theory.

<sup>13</sup>See Field (2009) for an interesting discussion for the relationship between logic and rational constraints on doxastic attitudes, with a focus on non-classical logics.

attribution to an agent of credences that violate Negation Coherence would never be an optimal way of making sense of that agent's behavior. However, it remains unclear what sort of argument could show that violations of Negation Coherence are impossible without also showing that violations of coherence, *simpliciter*, are impossible. Stalnaker (1984), for instance, defends a view of propositional attitudes where what it is for an agent to have some set of propositional attitudes is for the attribution of those propositional attitudes to be part of an optimal explanation and rationalization of that agent's behavior.<sup>14</sup> Stalnaker argues that on this view, it is impossible to have inconsistent beliefs or to fail to believe a necessary truth, and the same arguments would support the claim that it is impossible to have probabilistically incoherent credences.<sup>15</sup> So, someone like Stalnaker might argue that it is impossible to violate Negation Coherence, but only on grounds that suggest that it is impossible to violate coherence in any way, thus making the DBA wholly superfluous.

My purpose is not to adjudicate between these two positions. The debate over the proper theory of propositional attitudes is long-running, and I have nothing new to add to it here. I merely want to emphasize that there is some intuitive pull toward saying that violations of Negation Coherence are possible, even if certain theories of propositional attitudes say otherwise. Moreover, it is difficult to see how one might argue that violations of Negation Coherence are impossible, while maintaining that it is possible to violate coherence in other ways. In this way, I doubt whether it is possible to argue that violations of Negation Coherence are impossible without leaving the DBA with no work left to do.

#### 4.7.2 Irrational to Violate Negation Coherence?

Even if violations of Negation Coherence are *possible*, it might be that they are irrational, albeit for reasons wholly independent of the DBA. Now, there

---

<sup>14</sup>This Radical Interpretation theory of propositional attitudes is also defended by Davidson (1973) and Lewis (1974).

<sup>15</sup>Stalnaker treats purported ignorance of necessary truths as a sort of metalinguistic ignorance, or ignorance of which proposition is expressed by a given sentence. And he treats alleged cases of inconsistent beliefs as involving *fragmentation* - the agent is represented as having multiple belief states, each of which is consistent and which govern the agent's actions in different behavioral situations. The fact that different belief states are operative in different behavioral situations gives the misleading appearance of inconsistent beliefs.

are many arguments that have been advanced which entail that Negation Coherence is irrational, but they do so only by entailing that all forms of incoherence are irrational, and thus serve to replace, rather than supplement, the DBA.

So, for instance, the Representation Theorem Argument, defended by Maher (1993), among others, says that it is a requirement of rationality that your preferences have a certain structure, and that if your preferences have that structure, then you are best represented as an expected-utility-maximizer with probabilistically coherent credences. Hence, if you are rational, then you will have coherent credences.

Scoring Rule Arguments, discussed by Joyce (1998), hold that for every incoherent credence function, there is some coherent credence function which scores better on some measure of expected accuracy. Together with an argument that it is rational to have credences which maximize this sort of expected accuracy, Scoring Rule Arguments say that it is irrational to have incoherent credences, since each incoherent credence function is rationally inferior (in virtue of having lower expected accuracy) than some coherent credence function.

While I do not find these arguments convincing, I do not want to get into the details of these arguments here, for they have already generated an extensive literature.<sup>16</sup> The important thing to note is just that these arguments purport to show not that Negation Coherence in particular is a requirement of rationality, but that full probabilistic coherence is a requirement of rationality. In this way, if you are convinced by one of these arguments, then you will think that the DBA is just superfluous.

Might there be some argument that only shows that violations of Negation Coherence are irrational, thus leaving the DBA with an important role to play in getting us from Negation Coherence to full coherence? I doubt it. I grant that violations of Negation Coherence often seem more bizarre than some other sorts of incoherence, since Negation Coherence is a constraint that applies only to pairs of propositions. In this way, it seems that it should be particularly easy to satisfy Negation Coherence, in just the way that it is easier to avoid blatantly contradictory beliefs (like believing both  $P$  and  $\neg P$ ) than to avoid more subtly contradictory beliefs. But this consideration does not seem like a good argument that all violations of Negation Coherence

---

<sup>16</sup>See especially Meacham and Weisberg (forthcoming) for criticism of Representation Theorem Arguments and Gibbard (2008) for criticism of Scoring Rule Arguments.

are irrational. For one thing, you might violate Negation Coherence for principled reasons rather than due to sloppy thinking, at least if adherence to certain non-classical logics like intuitionism and paraconsistent logic means that you really do violate Negation Coherence. For another, the mere fact that Negation Coherence is relatively easy to satisfy doesn't mean that it is irrational to violate it. There are lots of things that are easy to do but are not rationally required.<sup>17</sup>

Of course, it remains possible that it is just a brute, unexplained requirement of rationality that you satisfy Negation Coherence. This would be somewhat unsatisfying, but unsatisfactoriness doesn't entail falsehood. Chains of explanations must come to an end somewhere, and perhaps the explanation of why it is irrational to have incoherent credences ends in the brute fact that it is irrational to violate Negation Coherence. I think that this is the position that must be adopted in order to give the DBA a role to play in showing that all rational agents satisfy probabilistic coherence. If we don't rule out violations of Negation Coherence as impossible or irrational, it is possible to have incoherent credences that cannot be shown by the DBA

---

<sup>17</sup>Actually, it may be possible to argue that violations of Finite Additivity in particular are irrational. And, since violating Negation Coherence entails violating either Finite Additivity, Normalization, or both, this claim would mean that some violations of Negation Coherence - those that stem from violations of Finite Additivity - are irrational. The idea is that violations of Finite Additivity yield inconsistent calculations of expected utility, and hence inconsistent behavioral commitments. Suppose  $A$ ,  $B$ , and  $C$  are mutually exclusive and jointly exhaustive. You assign credence 0.2 to  $A$ , 0.2 to  $B$ , and 0.2 to  $C$ , but you assign credence 0.8 to  $B \vee C$ , thus violating Finite Additivity. Consider an act  $\varphi$  that yields utility 3 if  $A$  and utility  $-1$  if either  $B$  or  $C$  is true. One way of calculating expected utility takes each of  $A$ ,  $B$ , and  $C$  to be an outcome, so that the expected utility of  $\varphi$  is  $0.2 \times 3 + 0.2 \times -1 + 0.2 \times -1 = 0.2$ . And so  $\varphi$  looks like a good thing to do. But we could also take  $B \vee C$  to be an outcome, since  $\varphi$  yields the same utility whichever disjunct is true, and therefore use your credence in the disjunction  $B \vee C$  in the calculation. And so the expected utility of  $\varphi$  is  $0.2 \times 3 + 0.8 \times -1 = -0.2$ . On this way of doing things,  $\varphi$  looks like a bad thing to do. Since it should be equally legitimate to think of outcomes in a maximally fine-grained way (as on the first way of calculating expected utilities) and to think of them in a more coarse-grained way, lumping together propositions which are assigned the same utility (as on the second way of doing things), you could interpret this as a situation in which your credences commit you both to  $\varphi$ -ing (since  $\varphi$  has positive expected utility on one way of calculating expected utilities) and to not  $\varphi$ -ing (since it has negative expected utility on another way of calculating expected utilities). Hence, violations of Finite Additivity yield inconsistent behavioral commitments, and are therefore irrational. Still, this argument would not show that violations of Negation Coherence that stem from violations of Normalization are irrational.



to be irrational. But if we rule out violations of Negation Coherence as impossible or irrational only by ruling out *all* violations of coherence as impossible or irrational, then the DBA is superfluous. To maintain, then, that the DBA has an important role to play in showing that all rational agents satisfy coherence, we are forced to hold that it is just a brute requirement of rationality that you satisfy Negation Coherence.<sup>18</sup>

## 4.8 Conclusion

I have argued that if we think of credences as licensing actions only in virtue of expected utility considerations, then the DBA cannot be used to condemn violations of Negation Coherence as irrational. Worse, it is actually possible for incoherent credences to fail to license accepting any Dutch Books, provided those credences violate Negation Coherence. Hence, the DBA can at best show that, once one satisfies Negation Coherence, any deviation from probabilistic coherence is irrational. In this way, the DBA is a partial result and falls short of showing that all incoherent credences are irrational.

Then, unless we are prepared to accept that it can in some cases be rationally permissible to have incoherent credences, we face a dilemma. First horn: It is impossible to be incoherent in any way. Reasons for thinking that it is impossible to violate Negation Coherence generalize to suggest that it is impossible to violate coherence at all. Second horn: Some structural constraints on credences (in particular Negation Coherence) are brute requirements of rationality that must be accepted for their intuitive, pretheoretic plausibility rather than on the basis of any argument. No independent arguments for the irrationality of violating Negation Coherence are compelling.<sup>19</sup>

---

<sup>18</sup>Perhaps accepting Negation Coherence, and possibly even all of probabilistic coherence, as a brute requirement of rationality would not be so bad after all. Compare the requirement that binary beliefs be logically consistent. We have no Dutch Book-style exploitability argument for this requirement of rationality. You might think that beliefs must be logically consistent, because if your beliefs are inconsistent, you are guaranteed to have at least one false belief, and that is bad. But this is not a compelling argument, since if your beliefs are inconsistent (and closed under entailment), you are also guaranteed to have at least one true belief, and that is good. So, I doubt whether there is any compelling argument for the requirement that beliefs be consistent. Instead, it is a pretheoretically compelling requirement of rationality. I propose that we should think of Negation Coherence in the same way.

<sup>19</sup>This assumes, of course, the failure of Representation Theorem and Scoring Rule

Therefore, provided that it is so much as *possible* to have incoherent credences, some structural constraints on credences must be brute rational requirements.

---

Arguments. Again, I reject these arguments for reasons given in Meacham and Weisberg (forthcoming) and Gibbard (2008), which cannot be recounted here for reasons of space.

# Bibliography

- Anscombe, Elizabeth. *Intention*. Oxford: Basil Blackwell, 1957.
- Anscombe, G.E.M. *Intention*. Oxford University Press, 1957.
- Arntzenius, Frank, Adam Elga, and John Hawthorne. “Bayesianism, Infinite Decisions, and Binding.” *Mind* 113 (2004): 251–283.
- Bermudez, Jose Luiz. *Decision Theory and Rationality*. Oxford University Press, 2009.
- Bratman, Michael. *Intentions, Plans, and Practical Reason*. CSLI, 1987.
- Bratman, Michael. “Reflection, Planning, and Temporally Extended Agency.” *Structures of Agency*. . Oxford University Press, 2007.
- Buchak, Lara. “Review of José Luis Bermúdez, *Decision Theory and Rationality*.” *Notre Dame Philosophical Reviews* 2009 (2009).
- Chisholm, Roderick. “Contrary-to-Duty Imperatives and Deontic Logic.” *Analysis* 24 (1963): 33–36.
- Christensen, David. “Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers.” *Journal of Philosophy* 93 (1996): 450–479.
- Davidson, Donald. “Radical Interpretation.” *Dialectica* 27 (1973): 314–328.
- Davidson, Donald, J. C. C. McKinsey, and Patrick Suppes. “Outlines of a Formal Theory of Value, I.” *Philosophy of Science* 22 (1955): 140–160.
- Finetti, Brunode . “La Prevision: ses lois logiques, ses sources subjectives.” *Annales de l’Institut Henri Poincar* (1937).

- Dougherty, Tom. "On Whether to Prefer Pain to Pass." *Ethics* 121 (2011): 521–537.
- Dougherty, Tom. "A Deluxe Money Pump Argument." (ms).
- Elga, Adam. "Subjective Probabilities Should be Sharp." *Philosopher's Imprint* 10 (2010).
- Field, Hartry. "What is the Normative Role of Logic?." *Aristotelian Society Supplementary* 83 (2009): 251–268.
- Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66 (1969): 829–839.
- Gibbard, Allan. "Rational Credence and the Value of Truth." *Oxford Studies in Epistemology, Vol 2*. . Oxford University Press, 2008.
- Hajek, Alan. "Dutch Book Arguments." *The Oxford Handbook of Rational and Social Choice*. . Oxford University Press, 2008.
- Hare, Caspar. "Take the Sugar." *Analysis* 70 (2010): 237–247.
- Hinchman, Edward. "Trust and Diachronic Agency." *Nous* 37 (2003): 25–51.
- Howson, Colin and Peter Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing Company, 1993.
- Jackson, Frank and Robert Pargetter. "Oughts, Options, and Actualism." *Philosophical Review* 95 (1986): 233–255.
- Jeffrey, Richard. *The Logic of Decision*. University of Chicago Press, 1965.
- Jeffrey, Richard. "Preference among Preferences." *Probability and the Art of Judgment*. . Cambridge University Press, 1992.
- Joyce, James. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65 (1998): 575–603.
- Joyce, James. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.
- Joyce, James. "How Probabilities Reflect Evidence." *Philosophical Perspectives* 19 (2005): 153–178.

- Kavka, Gregory. "The Toxin Puzzle." *Analysis* 43 (1983): 33–36.
- Kemeny, John. "Fair Bets and Inductive Probabilities." *Journal of Symbolic Logic* 20 (1955): 263–273.
- Kolmogorov, Andrey. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer, 1933.
- Kolodny, Niko. "IX-How Does Coherence Matter?." *Proceedings of the Aristotelian Society* 107 (2007): 229–263.
- Kolodny, Niko. "Why Be Disposed to Be Coherent?." *Ethics* 118 (2008): 437–463.
- Korsgaard, Christine. *The Constitution of Agency*. Oxford University Press, 2008.
- Lehman, R. Sherman. "On Confirmation and Rational Betting." *Journal of Symbolic Logic* 20 (1955): 251–262.
- Levi, Isaac. "On Indeterminate Probabilities." *Journal of Philosophy* 71 (1974): 391–418.
- Lewis, David. "Radical Interpretation." *Synthese* 27 (1974): 331–344.
- Lewis, David. "Causal Decision Theory." *Australasian Journal of Philosophy* 59 (1981): 5–30.
- Lewis, David. *Papers in Metaphysics and Epistemology*. Cambridge University Press, 1999.
- Lewis, David. "Why Conditionalize?." *Papers in Metaphysics and Epistemology*. Cambridge University Press, 1999.
- Maher, Patrick. *Betting on Theories*. Cambridge University Press, 1993.
- Meacham, Christopher and Jonathan Weisberg. "Representation Theorems and the Foundations of Decision Theory." *Australasian Journal of Philosophy* (forthcoming).
- Parfit, Derek. *Reasons and Persons*. Oxford University Press, 1986.

- Pollock, John. "Rational Choice and Action Omnipotence." *Philosophical Review* 111 (2002): 1–23.
- Priest, Graham. "The Logic of Paradox." *Journal of Philosophical Logic* 8 (1979): 219–241.
- Raiffa, Howard. *Decision Analysis*. Wesley-Addison, 1968.
- Skyrms, Brian. "Dynamic Coherence and Probability Kinematics." *Philosophy of Science* 54 (1987): 1–20.
- Smith, Matthew Noah. "Practical Imagination and its Limits." *Philosophers' Imprint* 10 (2010).
- Stalnaker, Robert. *Inquiry*. The MIT Press, 1984.
- Taylor, Shelley and Jonathon Brown. "Illusion and Well-Being: A Social-Psychological Perspective on Mental Health." *Psychological Bulletin* 103 (1988): 193–210.
- Teller, Paul. "Conditionalization and Observation." *Synthese* 26 (1973): 218–258.
- Thomson, Judith Jarvis. *Rights, Restitution, and Risk*. Harvard University Press, 1986.
- Fraassen, Basvan . "Belief and the Will." *Journal of Philosophy* 81 (1984): 235–256.
- Wright, G.H.von . "Deontic Logic." *Mind* 60 (1951): 1–15.
- White, Roger. "Evidential Symmetry and Mushy Credence." *Oxford Studies in Epistemology, vol 3* . Oxford University Press, 2009.
- Williams, Bernard. *Moral Luck*. Cambridge University Press, 1982.
- Williamson, Timothy. *Knowledge and its Limits*. Oxford University Press, 2000.