Options for Decision Theory

Gary James Mullen

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

University of Leeds

School of Philosophy, Religion and the History of Science

August 2018

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

Above all, I'd like to thank my supervisors, Robbie Williams and Ed Elliott. Robbie's guidance over the past four years has helped me in every chapter, and the thesis wouldn't have been possible without him. Ed has been my supervisor for the past twelve months, and my work has benefitted greatly from his comments.

A big thank you to Adina Covaci, Emily Paul, Bryan Ross, and Alison Toop for making PhD life more fun.

And finally, a special thank you to Adina Covaci.

I dedicate this thesis to my family.

Abstract

Decision theory says that an agent ought to choose an option that is evaluated best in light of the agent's beliefs and desires. But what gets to count as an *option*? I propose an account of options, which, in a typical case, entails that your options are decisions that you are certain you can make. Although we normally talk as if actions are rational and irrational, it is the evaluations of *decisions* that are decisive in determining what an agent ought to do. Another upshot of this account is that options are determined from the agent's perspective, much like decision theoretic evaluations of options.

The motivations for my account are twofold. First, a puzzle: an option must be *available* to the agent on both a subjective and objective reading, but it looks difficult for a candidate option to be both. I say that options are determined from the agent's perspective because that is the best way to resolve this puzzle. Key to my argument will be a sophisticated formulation of decision theory, on which, it says that an agent ought to do *as much as she can of* the best option. The second motivation is that an account of options must deliver plausible verdicts in some tricky test cases. I say that options are decisions because this offers the best hope of doing that.

As well as construing options as decisions, I construe them as *counterfactuals* so that I can deal with otherwise puzzling cases where the agent is uncertain about her decision-making abilities. This is a novel construal of an option, one on which an option isn't action-like.

Finally, I look at the consequences of my account for a theory of rational deliberation. If options are decisions, it's natural to think that rational deliberation involves *decision instability*.

Table of Contents

Chapter 1 – Setting the Scene

1. Introduction

According to many theories, you ought to choose what is best *out of your options*. Brokering world peace would be great, but if it is not an option, then it is not something you ought to choose. Similarly, diverting a runaway trolley onto a helpless bystander is pretty bad, but if it is best out of a set of options (perhaps because the only other option is letting the runaway trolley career into a crowd), then it is something that you ought to choose. In short, when determining what you ought to do, options matter. This thesis is about options. In particular, it is about options *in decision theory*.

Decision theory is a theory about what an agent ought *rationally* to choose (as opposed to, for instance, what an agent ought *morally* to choose). Roughly, decision theory recommends that the agent evaluate an option by envisaging the possible outcomes of realising that option and evaluating how desirable those outcomes are. Ultimately, it says that the agent ought to choose the option that is evaluated best by this method.

So the question I'll be concerned with in this thesis is: what is an option for the purposes of decision theory? In the remainder of this chapter I'll do three things. First, I'll outline decision theory in more detail. Second, I'll present some assumptions I'll be making in this thesis. Finally, I'll preview what's to come.

2. What is Decision Theory?

Roughly, decision theory recommends that the agent evaluate an option by envisaging the possible outcomes of realising that option and evaluating how desirable those outcomes are. Ultimately, it says that the agent ought to choose the option that is evaluated best by this method. More precisely, decision theory says that the agent ought to choose the option with the greatest *expected utility* (*EU*). This will take some time to explain.

First of all, decision theory is a theory about what an agent ought to do *when facing a choice* or, as I'll sometimes put it, when she confronts a *decision problem*. At least on first pass, we're talking about all sorts of choices here: choosing what to wear in the morning, choosing what career to pursue, choosing whether to make a decision now or later. For instance, suppose you are going to a dinner party and you are to bring the wine. You remember that your hosts will serve either chicken or beef, but you don't remember which, though you think it more likely that they'll serve chicken. Moreover, you think chicken goes well with white, but beef goes well with red. You have no way of contacting the hosts and can only bring one bottle of wine. What do you do?[1]

Your decision problem here can be separated into *options*, *states,* and *outcomes*. The *options* are, very roughly, the actions available to you. Although there are cases where it's far from obvious what an agent's options are, in this case it appears that your options are "bring bottle of white wine" and "bring bottle of red wine". Often, the results of realising an option depend on external factors about which the agent is uncertain. These external factors are the *states*. In your decision problem, you are uncertain about whether your hosts will cook chicken or beef, and which they cook determines the results of your action. Finally, the outcomes are the results of realising a given option in a given state. For instance, bringing red given that beef is cooked results in the outcome of bringing the *right* wine, whereas bringing white given that beef is cooked results in the outcome of bringing the *wrong* wine. This information can be summarised in a *decision matrix*:

	Chicken	Beef
Red	Wrong wine	Right wine
White	Right wine	Wrong wine

[1] This example is from Jeffrey (1983, ch.1).

To obtain an option's EU, we have to consider two factors. First, we consider the agent's degree of belief for each outcome on the assumption that she realises each option. I'll assume that the agent's degrees of belief can be represented numerically by real numbers in the interval 0≤x≤1 – the higher the number, the greater the agent's degree of belief. For instance, perhaps your degree of belief that you will bring the wrong wine on the assumption that you bring red wine is 0.6 whilst your degree of belief that you will bring the right wine on the assumption that you bring red is 0.4. I'll use "degree of belief" and "credence" interchangeably. And "Cr(P)" will label the agent's credence in P. Your credences for each outcome on the assumption that you realise each option can be summarised in a *probability matrix*:

	Chicken	Beef
Red	0.6	0.4
White	0.6	0.4

The second factor we need to consider is the agent's degree of desire for the outcomes. I'll assume that the agent's degrees of desire can be represented numerically – the higher the number, the greater the agent's desire. For instance, perhaps you have a degree of desire of 10 for bringing the right wine and a degree of desire of 0 for bringing the wrong wine. I'll use "degree of desire" and "utility" interchangeably. And "U(P)" will label the agent's utility for P. Your utilities for each outcome can be summarised in a *utility matrix*:

	Chicken	Beef
Red	0	10
White	10	0

The EU of an option is obtained in two stages. First, we multiply the corresponding entries of the probability and utility matrices. The result in your case is:

	Chicken	Beef
Red	0	4
White	6	0

Second, we add the entries in each row. So the EU bringing red is 4 and the EU of bringing white is 6. Decision theory says that an agent ought to choose the option with greatest EU. So decision theory entails that you ought to choose white wine. (If two or more options are tied for EU, then decision theory says that it is permissible for the agent to choose any such option.)

Another way to think of the EU of an option A is that it is the average of the utilities of A's possible outcomes O, weighted by the agent's credence for O on the assumption that she performs A. In symbols, the EU of A is:

$$\sum_{O} Cr(O\|A)U(O)$$

Here "O" ranges over possible outcomes of the agent's choice; "Cr(O‖A)" labels the agent's credence in O on the assumption that she performs A. There is some disagreement about how to understand Cr(O‖A). There are two main schools. *Evidential Decision Theory* (*EDT*) says that Cr(O‖A) should be understood as Cr(O-and-A)/Cr(A). This is often taken as a measure of the extent to which the agent thinks A provides evidence for O. So one way of putting EDT is that it evaluates an option by looking at the extent to which the agent thinks the option is *evidence* for desirable and undesirable outcomes. In contrast, *Causal Decision Theory* (*CDT*) says that Cr(O‖A) should be understood as the agent's credence in the counterfactual *if A were performed then O would obtain*. This is often taken as a measure of the extent to which the agent thinks A causes O. So one way of putting CDT is that it evaluates an option by looking at the extent to which the agent thinks the option causes desirable and undesirable outcomes.

Finally, what is it to *choose* an option? I am going to start with a very simple account of choice, namely, that an agent chooses an option when she realises that option. So on this account, an agent chooses to, for instance, bet on black, not when she *decides* or *intends* to bet on black, but when she actually bets on black. (I start with this account, and failing to see any reason to change it, I end with this account.)

Let me note a couple of distinctive features of decision theory as I understand it. First, the relevant beliefs and desires of the agent (for the purposes of the calculation of EU) are the beliefs and desires *before* she chooses an option. After she chooses, she may change her beliefs because she considers her decision to be evidence for certain states of affairs. But the relevant beliefs and desires are the beliefs and desires the agent has *before* the agent chooses.[2]

Second, I see decision theory as providing a constraint on the agent's credences and utilities on the one hand, and her choices on the other. An alternate conception of decision theory sees it as providing a constraint on the agent's credences and utilities on the one hand, and her preferences over a certain class of objects on the other. The constraint is, roughly, that the preferences ought to be ordered by the expected utilities of their objects relative to the credences and utilities. In Savage (1972) these objects are *acts*, arbitrary functions from states to outcomes, whilst in Jeffrey (1983) they are arbitrary propositions. If this is how one sees decision theory, then my thesis is about an additional norm which I consider implicit in the above picture of decision theory, namely, that an agent ought to choose an option such that no other option is preferred to it. And I assume that the class of objects, over which preferences are defined, contain the options. My thesis is about this additional norm and, in particular, about what an *option* is as appealed to in this norm. Note that both Savage's and Jeffrey's theories are mostly silent on this. On Savage's theory, the "acts" are arbitrary functions from states to outcomes so include entities that can't be considered options in any ordinary sense. On Jeffrey's theory, the relevant class of objects are arbitrary propositions so again can't be considered options in any ordinary sense. Nevertheless, I will consider my thesis as about decision theory understood not as a constraint on credences-utilities and *preferences*, but as a constraint on credences-utilities and *choices*.

[2] See Sobel (1990) for discussion of this distinction.

3. Preliminaries & Assumptions

In this section I'll outline some assumptions I'll be making in this thesis. First, let me deal with some preliminaries (for the thesis). I'll often say that a particular option is *best*. This means that it has greatest EU. I'll also say that a particular option is *maximal*. This means that no other option has better EU. So two or more options might be *maximal*, but only one option will ever be *best*. I will use "A" and "B" to label arbitrary actions. Other conventions and terminological choices I'll introduce as I go along.

Now for the assumptions that I'll be making. I make these to simplify discussion. First, I'll go through some assumptions about agency. I'm construing *action* broadly to include three sorts of action. First, *outer actions,* i.e. actions that involve the agent interacting with the world, e.g. the pulling of a lever. Second, *decisions*, which I consider to be the tokening of a mental state that is directed upon an outer action and which typically guides the performance of the outer action. Third, *tryings*, which consists in the decision (or rather the mental state tokened in a decision) guiding the agent in her completion of the outer action. This broad construal of *action* is important because there is a natural sense in which a decision is not an action – however, I will be thinking of decisions as a type of action.

A very simple model of agency falls out from the above: a decision tokens a mental state directed towards an outer action, which is then involved in making the outer action happen, and the latter constitutes a trying to do that outer action. Although this is a toy model of what happens when we act, I don't think anything substantial hinges on it, and it simplifies the discussion to come.

This conception of the relation between an intentional mental state, trying, and outer action finds more precise expression in Adams & Mele (1992). According to their account, "trying to A is an event or process that has A-ing as a goal and is initiated and (normally) sustained by a pertinent intention" (1992, p.326). For instance, suppose Mary, unbeknownst to Benny, straps down Benny's right arm and subsequently tells him to raise it. A moment later, Benny is surprised that his right arm isn't in the air. According to their account, here's what happened: Benny intended to raise his arm and did stuff as a result of that intention. In particular, his nerves from his brain to his right arm fired and his muscles in his right arm contracted somewhat. Benny's doings – the nerve firings and the muscle contractions – *are* his trying to raise his right arm. That's because Benny's nerve firings and muscle contractions have raising his arm as a goal, and they are triggered and sustained by his intention to raise his arm. Trying, on this view, is the mediator between intention and outer action. Intentions trigger and sustain tryings, which lead to the outer action. I am sympathetic to this way of fleshing out my toy model of agency – except, of course, that I talk of a *decision* rather than an *intention*.

Now I turn to assumptions about agency with respect to time. I see decision theory as generating an obligation on an agent at a particular time, namely, *at the time she faces her choice*. This is the time before she's done anything. The time at which the agent might ask herself "what ought I do?". The obligation consists in an obligation to realise a certain option.

I assume an option is an action.[3] I'm thinking of an action (and hence an option) as corresponding to a specific time – this is the time at or over which the action takes place. I am also assuming that the relevant actions are *immediately performable*. That is, I assume that they start (and possibly end) immediately after the time the agent faces her choice. Let t_i be this time.

When I say that an action takes place *at a time*, then I mean that it starts and ends at that time. I call such an action a *minimal action*. Given the assumption that options are immediately performable, minimal actions start and end at t_i, that is, they are *at t_i*. In addition to minimal actions, there are also *extended actions*. Extended actions take place over a specific time period. They last longer than minimal actions. Given that options are immediately performable, they take place over some time period starting with t_i.

[3] More precisely, I make this assumption *for now*. I will argue in Chapter 4 that this isn't *always* the case: sometimes, options are counterfactuals.

For instance, a typical minimal action is raising one's arm, because this starts and ends at a single time. If an agent has raising her arm as an option, then strictly-speaking, it will be *raising her arm at* t_i that is the option. A typical extended action is raising one's arm five times consecutively, because this is completed over a time period. If an agent has this action as an option, then strictly-speaking, it will be *raising her arm from* t_i *to t'* (where t' is some future time) that is the option.

Now I turn to assumptions about abilities. Options obviously have something to do with the agent's *abilities* – what she *can* do. (I will use "agent is able to A" and "agent can A" interchangeably.) But there appear to be many sorts of abilities. For instance, when we say that Sally can't speak French we might mean that she can't speak French in the sense that she's never learnt French; we might mean that she can't speak French in the sense that she's currently drunk or asleep; we might mean that she can't speak French in the sense that there's no amount of training that she could receive that would make her speak French. So which sort of 'can' is relevant for my purposes?

There is a distinction between general and specific abilities. As a rough approximation, a general ability is an ability an agent has in a wide range of circumstances. It involves having a competence and so depends on what the agent is internally like. For instance, Andy Murray has a general ability to make a tennis serve. He has this competence, through his hours of training. In contrast, a specific ability is an ability an agent has in a particular circumstance at a particular time. It involves having "what it takes" and cooperative circumstances at that time. For instance, Andy Murray has a specific ability to make a tennis serve when he's on court, with a racquet etc. but he doesn't otherwise.

An agent might have a general ability to perform some action A without having the specific ability to A. For instance, suppose Andy Murray is not on a tennis court. He has the general ability to serve – through his hours of practice – but not the specific ability to serve – because he's not on a tennis court. More controversially, an agent might have a specific ability without having a general ability. For instance, I played tennis last weekend and made a serve despite having no competence in this (it was sheer fluke that I made a serve). I didn't have a general ability to serve and yet you might think that I had the specific ability to serve – after all, I did in fact make a serve.[4]

It's specific abilities that are of interest to the question of options. If Andy Murray knows that he has a general ability to serve but also knows that he is nowhere near a tennis court (so doesn't have a specific ability), then clearly *serving* doesn't count as an option for him.

There are other agentive modalities (aside from specific and general abilities), but it's even more obvious that these are irrelevant to the question of options, so I put them aside. For instance, there are *potentialities*. I have the *potential* to speak Mandarin (even though I don't have that specific ability nor that general ability), whereas a lion does not.[5]

From now on when I talk of *ability* I will mean a *specific ability*, unless stated otherwise. Can we be a bit more specific about what it is to have a (specific) ability to perform some action A? At a minimum, it looks like it includes:

- The agent is physically able to A: there is no physical internal condition that is stopping the agent from A-ing. For instance, if A involves moving limbs, then the agent is not paralyzed.
- The agent is psychologically able to A: there is no mental internal condition that is stopping the agent from A-ing. For instance, the agent is not in a state of unconsciousness.
- The agent is relevantly situated. For instance, if A involves doing something to an object, then the agent is in the vicinity of the object rather than, for instance, in another country.
- The agent's environment cooperates. For instance, if A involves moving the agent's limbs, there is no freak gust of wind that would render the agent incapable of controlling her limbs.[6]

[4] For the distinction between general and specific abilities, see Whittle (2010); Maier (2015); Glick (2012); Mele (2003).

[5] This example is from Maier (2015).

[6] These four conditions are from Haji (2002, p16).

These conditions correspond roughly to *opportunity*. A notion of *ability* that makes these four conditions individually necessary and jointly sufficient is often called a *simple ability* (see Mele 2003 and Zimmerman 2008, pp.132-3). This sort of ability isn't strong enough. For suppose that I have a die in my hand that I'm about to roll. I have a simple ability to roll a six on the die. That's because there is no mental or physical internal condition stopping me, so I am psychological and physically able to roll a six; I have the die in my hand so I am relevantly situated; there is as much chance as it landing on a six as any other number, so the environment cooperates. This is clearly not the relevant sort of ability. I know that I can roll a six (in this weak sense of ability), but rolling a six is not an option for me. (Similarly, suppose that there is a locked safe in front of me and that I have no idea what the combination is. I have a simple ability to open the safe. Again, this is clearly not the relevant sort of ability. I know that I can open the safe in this weak sense of ability, but it clearly isn't an option for me.)

There are a couple of ways of constructing a stronger notion of *ability*. Nothing in this thesis will depend on a particular way of fleshing this out, so I'll be brief. The two ways can be brought out by different reactions to Austin's golfer. Austin (1956, p.218) considers a skilled golfer who tries to sink a putt but misses. Did the golfer have the ability to sink the putt? Here intuitions diverge.

One reaction is to say that the golfer's miss, considered in isolation, shows very little. The golfer had the ability to sink the putt even though his attempt ended in failure. You might elaborate on this by saying that the golfer had the ability to sink the putt because he possessed both the opportunity (embodied in the four conditions above) and a *general ability* to sink the putt. A general ability to sink the putt entails a certain reliability with sinking the putt. Roughly, an agent has a general ability to sink the putt only if a suitable proportion of her tryings to sink the putt (in circumstance of the same type as her actual circumstances) would be successful. This is compatible with the occasional misfiring – and this is what happened in the golfer's case. As required, this account entails that I don't have an ability to roll a six when I have the die in my hand – because too few of my tryings to roll a six would end in success in similar circumstances.[7]

An alternative reaction is to say that the golfer *can't* sink the putt – his attempting and failing shows precisely this. This is the reaction of Honoré (1964) and is discussed in Portmore (2017, ch.3) as *Tryism*. This suggests an alternative approach according to which there is no entailed reliability condition on having an ability to perform some action A. The account might be something like this: an agent has the ability to A (in a certain circumstance) iff the agent is such that if she were to try to A (in that exact circumstance), then she would A. Suppose again that I have a die in my hand. This account says that I have an ability to roll a six depending on how we flesh out the case. If, for instance, I would in fact roll a five, then I can't roll a six – because my trying to roll a six wouldn't be successful. However, if I would in fact roll a six then I can roll a six – because my trying to roll a six would be successful. This is not immediately a problem: if the agent knows that she can roll a six in this sense, i.e. she knows that her trying to roll a six would lead to her rolling a six, then arguably *rolling a six* is an option.

The difference between these two approaches is that the first makes the ability more robust in that it depends on what the agent is like (and does) in circumstances other than the one she's currently in. That's not surprising, because what it takes to have a (specific) ability, on that approach, is to have a general ability. Roughly, a (specific) ability, on this approach, is a general ability plus opportunity. The second, in contrast, makes a (specific) ability depend only on the agent's current circumstances.

So to wrap up the discussion of abilities, it is *specific abilities* (as opposed to general abilities or potentialities) that are of interest to the decision theorist. An agent has a specific ability to perform some action A just when she has the opportunity to A (i.e. is physically and psychologically able, is relevantly situated, and the environment cooperates) and either (i) has a certain reliability with A-ing in

[7] This sort of account of "can" is offered by Mele (2003), which is his intentional "can" or I-ability; Wedgwood (2013) also offers such an account.

similar circumstances or (ii) would A if she were to try to A in those circumstances. The difference between (i) and (ii) won't matter for my purposes.

4. Preview of Thesis

Now I turn to previewing what's to come. This thesis is about options in decision theory. To the best of my knowledge, Weirich (1983), Sobel (1983), Pollock (2002), Hedden (2012) and Schwarz (2017) are the only works with substantial discussions on the topic of *decision theoretic* options. There is substantial discussion of options *for moral theories*.[8] However, this is conducted in a framework that's different in a number of important ways.

First, the moral options literature mostly assumes an objective evaluation of options whereas decision theory uses a subjective evaluation of options, that is, an evaluation from the agent's perspective. This is important because if options are evaluated objectively, then there is little appeal to subjective constraints on options. Subjective constraints on options will play an essential role in each chapter of this thesis.

Second, as Jackson & Pargetter (1986) put it, there are two questions we can ask when we talk about what an agent ought to do. First, we can ask, of a particular action, ought an agent do it? Second, we can ask, at or over a particular time, what ought an agent do? It's the second question that interests the decision theorist. However, the moral options literature is mostly interested in the first question.

Third, unlike in the moral case, in decision theory there are (reasonably) precise methods to evaluate an option. There is a dispute between evidential and causal methods by which options are evaluated. But at least these two methods of evaluation are reasonably clear. This is vital for determining which prescriptions an account of options entails, something I will make use of throughout this thesis. So

[8] In particular, I'm thinking of the discussion about options for Utilitarianism, see Bergström (1968) and Prawitz (1968) (1970) for the start of this discussion. And I'm thinking of discussions of options in the Actualism vs Possibilism debate (see Jackson & Pargetter 1986).

the moral options literature is importantly different to the decision theoretic literature. Thus, in looking at options *for decision theory* this thesis charts (relatively) new territory.

Nevertheless, there are elements of the moral options literature that are relevant. There is a relevant discussion about whether options are pairwise incompatible (see Bergstrom (1968) (1971) and Prawitz (1968) (1970)). I will appeal to this discussion in Chapter 3.

There is also a lesson from the moral literature which puts my thesis into context. The lesson is that a suitable account of options might be relative to the *role* the theory of the *ought* is supposed to play. (In particular, is the theory supposed to be action-guiding, and if so, in what sense?)[9] This is even more true in the case of decision theory because there is a role decision theory allegedly plays over and above the roles moral theories allegedly play. Moral theories are said to play an evaluative role, that is, they serve as guides for our practice of praising and blaming agents. They are also said to play an action-guiding role: they serve as practical guides for an agent deliberating about what to do. Decision theory is alleged to play both of these roles plus a predictive-explanatory role: on the assumption that an agent is rational, we use decision theory to predict what she will do; and on the assumption that an agent who acted is rational, we explain what she did by citing that it was rational according to decision theory.[10]

I won't, however, be narrowing my focus to a particular role. Instead, I'll be looking at issues that cut across the different roles decision theory allegedly plays. I'll be looking at these issues and drawing conclusions about what an account of options must and can look like (relative to any role), rather than drawing conclusions about what they do look like relative to a specific role. As we'll see, strong

[9] See Timmerman & Cohen (2016) for different moral theories relative to different roles. And see Portmore (2017, ch.4) for an interpretation of this in terms of different accounts of options.

[10] See Hedden (2012) and Bermudez (2009) for more on these roles.

conclusions can be drawn without paying attention to the specific roles that decision theory allegedly plays.

After this introduction, the thesis is divided into four chapters corresponding to four issues. **Chapter 2** discusses a puzzle for accounts of options: the *Objective-Subjective Puzzle*. Decision theory evaluates options subjectively, that is, in light of the agent's beliefs and desires. For this reason, there are, in addition to the normal objective constraints on options, subjective constraints on options, that is, the agent must bear the correct epistemic relation to her options. This generates a puzzle, for it looks difficult to formulate an account of options that is sensitive to both subjective and objective features. This puzzle is discussed in Hedden (2012) and Schwarz (2017). I argue for the following account of options:

> **Subjective Actions**. Options are all and only the actions that the agent is certain she can do. (For some precisification of "actions".)

I explain how this satisfies the objective constraints on options if we adopt a slight reformulation of decision theory. As I initially glossed it, decision theory says that an agent ought to realise the best option. I propose to reformulate decision theory so that it says that an agent ought to *do as much a she can* of the best option – this is the *sophisticated formulation* of decision theory. The main consequence of this proposal is that options are determined from the inside – from the agent's perspective. An additional consequence is that this puzzle does not motivate a conception of options as inner actions (*pace* a number of authors including Hedden and Schwarz).

Nevertheless, I *do* think that options are decisions, but I think that this is motivated by concerns unrelated to the Objective-Subjective Puzzle. This is the subject of **Chapter 3**. I frame this chapter as follows: **Subjective Actions** is schematic in that "actions" awaits precisification; what is the correct precisification? I argue that the only way to generate sensible verdicts in concrete cases is to interpret "actions" as "decisions". So I argue for the following account of options:

> **Subjective Decisions**. Options are all and only the *decisions* that the agent is certain she can make.

This delivers sensible verdicts in concrete cases. In contrast, accounts of options that conceive of the options as outer actions fall prey to a number of difficulties. An argument of this form has been presented before (Sobel 1983). However, I reject this argument in favour of a new argument for **Subjective Decisions**.

Subjective Decisions has a number of important consequences. It does violence to our ordinary ways of speaking, it doesn't fit well with a popular picture of rational deliberation, it makes rational obligation *less* demanding than you might have thought, and it rules out an argument for the irrationality of performing a sequence of actions that leads predictably to disaster (e.g. in a money pump). Moreover, it makes the agent's predicted future irrationality always relevant, in contrast to some competing accounts. I will outline all of these consequences in more detail in Chapter 3.

Chapters 2 and 3 assume that the agent, for any decision, assigns an extremal credence (either 0 or 1) to her being able to make it. It is under this assumption that I offer **Subjective Decisions** as an account of options. I make this assumption to put aside a puzzle for any account of options. In **Chapter 4**, I explore this puzzle. The puzzle concerns cases where the agent is uncertain that she can *decide* on an action. For instance, suppose Brenda is uncertain that she can decide to ford a creek because an evil demon might strike her down just as she is about to decide on it. There appears to be no suitable ford-the-creek-like option. You might think that the ford-the-creek-like option is *decide to ford the creek*. This is a mistake because decision theory evaluates it in a way that wrongly ignores Brenda's doubts about being able to decide to ford the creek. There appears to be a missing option. This is the *Missing Option Puzzle*. A similar puzzle is discussed by Pollock (2002). I defend what I call the *Counterfactual Strategy*, which says that the ford-the-creek-like option is the counterfactual *if Brenda were able to decide to ford the creek, then she would decide to ford the creek*. This puzzle and solution are interesting in their own right but they also bear on the issues already discussed. First, I extend **Subjective Decisions** to generate counterfactual options in Brenda-type cases. Second, I explain how the Counterfactual Strategy pre-empts an objection to the conclusions reached in Chapters 2 and 3.

Let me flag now the conclusion of Chapters 2-4. This conclusion, unlike the discussion of chapters 2 and 3, does *not* assume anything about the agent's attitudes towards her decisions. I propose the following account of options:

> **Subjective Decisions with Counterfactuals.** An agent's options are all the decisions that the agent is certain she can make. Also, for any decision to perform some action A, if the agent is uncertain about being able to make that decision, the counterfactual *if the agent were able to decide on A, then she would decide on A* is also an option.

Along with this, I propose the sophisticated formulation of decision theory, on which, the agent ought to *do as much as she can of* the best option.

I call the resultant conception of decision theory, understood as including the account of options and the sophisticated formulation, *Sophisticated Subjective Decision Theory (SSDT).* This is sophisticated in two ways. First, in adopting a sophisticated formulation of decision theory. Second, in adopting a sophisticated variant of **Subjective Decisions** (namely, **Subjective Decisions with Counterfactuals**).

Finally, **Chapter 5** turns to rational deliberation. I say that SSDT leads to the *Sobellian picture of rational deliberation*. According to this picture, the agent makes a decision that is evaluated best; she subsequently becomes confident of executing that decision; she then re-evaluates her options with her updated credences, and again makes a decision that is evaluated best (see Sobel 1990). I present an objection to the Sobellian picture: a self-aware Sobellian deliberator is forced into epistemic irrationality. This is bad news for SSDT because SSDT leads to the Sobellian picture. So in this chapter I reject that objection to Sobellian deliberation.

Chapter 2 – The Objective-Subjective Puzzle

1. Introduction

There is a puzzle for any account of options. Roughly, the puzzle is that, on the one hand, an option must be such that the agent can perform it; on the other hand, an option must be such that the agent *believes* that she can perform it. This is puzzling because these constraints force the option set to be sensitive to objective *and* subjective features of the agent's decision problem, and it looks difficult to do both. I call this the *Objective-Subjective Puzzle*. In effect, this tension is a tension between two ways of looking at an agent's options. On one conception, they are determined from the inside – from the agent's perspective on her abilities. On the other conception, they are determined from the outside – by paying attention to the agent's actual abilities.

I recommend dropping the constraint that says an option must be such that the agent can do it. Or more precisely, I recommend *replacing* it with a weaker constraint, namely, that 'ought' implies 'can' (*OIC*). This opens up room for the following solution. Options are all and only actions that the agent is certain she can do (for some precisification of "actions"). I call this account of options **Subjective Actions**. This trivially satisfies the requirement that the agent believes she can perform each option. However, it looks like it violates OIC. In response, I propose that decision theory is formulated so that it says not "the agent ought to realise the best option" but "the agent ought to do as much she can of the best option". In cases where the agent's beliefs about her abilities are so out of kilter with reality that she can't realise any part of her best option, then there is nothing that the agent ought to do.

Decision theory's evaluation of an option is determined from the perspective of the agent – it is the agent's beliefs and desires that determine which option is best. The upshot of this chapter is that decision theoretic options are *also* determined from the inside.

In what follows, I'll present the puzzle and say why construing options as decisions doesn't resolve it (section 2), then I'll present my solution to it (section 3). In the remaining sections I'll elaborate on some crucial elements of the preceding discussion: construing options as decisions (section 4), and the objective and subjective constraints at the heart of the puzzle (sections 5 and 6). Finally, there are two appendices. Here, I compare my discussion to Hedden (2012) and Schwarz (2017), in which similar puzzles are discussed.

2. The Objective-Subjective Puzzle

2.1 The Objective Constraint

The puzzle is generated by two constraints on options. I'll motivate both quickly here and then come back to a proper defence of them later. The first constraint is the Objective Constraint, which says that if A is an option, then the agent can A. "A" here (and in what follows) ranges over actions, which (as said in Chapter 1) I am construing broadly to include both outer actions (e.g. pulling a lever) and inner actions (e.g. deciding to pull a lever).

The motivation for the Objective Constraint is that it ensures decision theory satisfies the principle that 'ought' implies 'can' (henceforth *OIC*).[11] In turn, the motivation for OIC is that it seems strange to ascribe to an agent an obligation that she can't fulfil. Even if I am certain that I can cure cancer and curing cancer would have very good consequences, if I can't cure cancer, then it is odd to think that I ought to cure cancer. It may be that I ought to *try* to cure cancer, but it is weird to say that I ought to cure cancer.[12] The Objective Constraint ensures OIC because decision theory says that the agent ought to realise the option that is best, so if options are actions that the agent can perform, then the agent will be able to realise the best option.

[11] See Hedden (2012) and Schwarz (2017) for similar motivations for the Objective Constraint.

[12] See Streumer (2007) and Andric (2017) for similar arguments for OIC.

2.2 The Certainty Constraint

The second constraint is the Certainty Constraint, which says that if A is an option, then the agent is certain that she can A. The motivation for the Certainty Constraint is as follows. Suppose that you're stood in front of two levers *a* and *b* that, when pulled, deliver $100. You're given the choice of pulling *one* of the levers. But here's the catch: sometimes lever *b* doesn't work – it sometimes jams, in which case you wouldn't be able to pull it. You assign 0.5 credence to lever *b* jamming. Which one do you go for? Obviously you pull lever *a*.

Suppose (for *reductio*) that *pull lever a* and *pull lever b* are your options here. *Pull lever a* receives the same EU as *pull lever b*, because both deliver $100. So decision theory says that it is permissible to pull lever *b*. However, it's impermissible for you to pull lever *b*. So *pull lever a* and *pull lever b* are not your options.

What's gone wrong here? Clearly there is a pull-lever-*a*-like option and a pull-lever-*b*-like option in this case. Decision theory misevaluates *pull lever b* because it focuses only on its possible outcomes and doesn't take into account your doubts about being able to do it. That means the option set {pull lever *a*, pull lever *b*} delivers the wrong verdict. Extrapolating from this case suggests that an agent must always be certain that she can perform each option, otherwise decision theory will misevaluate them. Hence the Certainty Constraint.[13]

Let me pre-empt a couple of worries about this motivation. First, you might think that *pull lever b* is not an option because you in fact can't do it. But note that I didn't make this assumption above. We can suppose that, in actual fact, you can pull lever *b*. What's doing the work in the above argument is merely that you assign some credence to being unable to pull lever *b*. Second, note that it is *certainty* that is required. For *pull lever b* is misevaluated no matter how small your doubts about being able to pull lever *b*. Even if you assign just 0.01 credence to lever *b* jamming, this is relevant, and means that you ought to pull lever *a*. Hence *certainty* is what's

[13] See Weirich (1983) (2004, p.26), Hedden (2012) and Pollock (2002) for similar motivations for the Certainty Constraint.

needed. (We will, however, see a qualification to the Certainty Constraint later, in section 6.)

2.3 Generating the Puzzle

That completes the two constraints on options. These constraints generate a puzzle, because it looks difficult to formulate a plausible account of options that satisfies both of them.

To see this, first consider the minimal account that satisfies the Certainty Constraint:

> **Subjective Actions**. Options are all and only the actions that the agent is certain that she can do.

"Actions" awaits precisification – perhaps as "outer actions", perhaps as "decisions". **Subjective Actions** struggles to satisfy the Objective Constraint because it seems possible for the agent to be certain that she can do something that she can't in fact do. For instance, I might be certain that I can make a 40 yard pass to a teammate on the other side of the pitch, but be mistaken about this.

Second, consider the minimal account that satisfies the Objective Constraint:

> **Objective Actions**. Options are all and only the actions that the agent can do.

Again, "actions" awaits precisification. **Objective Actions** struggles to satisfy the Certainty Constraint because it seems possible for the agent to be able to do something but for her to be uncertain about whether she can do it – indeed, it seems that she could be certain that she *can't* do it or she might lack beliefs altogether about it. For instance, I might be able to beat an opponent at chess but lack the belief that I can.

So whether we start with a minimal account satisfying the Certainty Constraint or a minimal account satisfying the Objective Constraint, it looks difficult to satisfy the other constraint. The root of the tension is that it's possible that the agent has mistaken beliefs about her abilities.

Now you might think that it's not difficult to formulate an account of options that satisfies the two constraints. For consider the following account:

> **Objective-Subjective Actions**. Options are all and only the actions that (i) the agent can do, and (ii) the agent is certain she can do.

(Again, "actions" awaits precisification.) This account effectively combines the two constraints to generate an account of options. It trivially satisfies the constraint, but it is implausible because it delivers intuitively wrong verdicts in cases where the agent is mistaken about her abilities.

Consider a case where there are three levers in front of you – *a*, *b*, and *c*. Suppose that you can pull *a*, that you can pull *b*, but that you *can't* pull *c*. Suppose you are certain of the following three things: that you *can't* pull *a*, that you can pull *b*, that you can pull c. Moreover, suppose you know that you can only pull *one* of these levers. Finally, suppose that pulling *a* and pulling *c* are great, but pulling *b* is terrible.

What does **Objective-Subjective Actions** entail in this case? It looks like it entails that you ought to pull *b*, because that is your only option – it is the only action that you are correctly certain that you can do. This is the wrong verdict. I think it's clear that there is *no* sense in which you ought to pull *b*. Pulling *b* is terrible, and there is both a distinct lever that you are certain you can pull which is subjectively better and a distinct lever which you can actually pull which is subjectively better.

In summary, **Objective-Subjective Actions** satisfies the constraints but is objectionable. Putting it aside, it's difficult to formulate an account of options that satisfies the constraints. **Subjective Actions** and **Objective Actions**, for instance, satisfy one constraint but not the other. Formulating a plausible account of options that satisfies the constraints is the Objective-Subjective Puzzle.

2.4 Why Decisions Don't Help

I've argued that **Subjective Actions**, **Objective Actions** and **Objective-Subjective Actions** all fail to satisfy the constraints. These accounts are schematic, in that "actions" awaits precisification. You might think that the argument doesn't go through if "actions" is interpreted as "decisions". You might think that the problem with those accounts stems from cases where the agent is mistaken about her abilities, but the agent is in a privileged epistemic position with respect to whether she can perform decisions. However, this is a wrong because there are agents who are mistaken about the decisions that they can perform just as there are agents who are mistaken about the outer actions that they can perform.

There are at least four circumstances in which I might be unable to make a decision. First, when there is some external force preventing me forming a decision – for instance, an evil demon ready to strike me down just as I am about to make a decision. Second, when I have a pathological psychology. For instance, if I'm addicted to smoking, then I might be unable to decide to stop smoking. Third, when I have some deep dislike of doing something. For instance, I might find the idea of torturing kittens so abhorrent that I can't decide to torture kittens. Fourth, when I have certain intentional mental states which preclude me from making certain decisions. For instance, suppose I believe that my decision to go on a diet would be completely ineffective. Then I might be unable to make that decision.

These four circumstances are all circumstances about which I might be mistaken, so I might be mistaken about my ability to make the corresponding decisions. This suffices to rule out decisional versions of **Subjective Actions**, **Objective Actions**, and **Objective-Subjective Actions** as solutions to the puzzle.

More precisely, first consider the decisional version of **Subjective Actions**:

> **Subjective Decisions**. Options are all and only the decisions that the agent is certain that she can make

This fails to satisfy the constraints because I might be certain that I can make a decision that I in fact can't make. For instance, I might be certain that I can decide to stop smoking but unbeknownst to me I'm addicted to smoking.[14] Second, consider **Objective Decisions** – options are all and only the decisions that the agent can *in fact* do. This fails because I might be able to make a decision that I am uncertain that I can make. For instance, I might assign some credence to there being an evil demon that would strike me down just as I'm about to decide to donate to charity, but in fact there isn't. Finally, consider **Objective-Subjective Decisions** – options are all and only the decisions that I am certain that I can make and that I can in fact make. This fails because, like **Objective-Subjective Actions**, it leads to a shortage of options in cases where the set of decisions-that-I am-certain-I-can-make is distinct from the set of decisions-that-I-can-in-fact-make. For instance, suppose there are three levers in front of you, *a*, *b*, and *c*. Suppose that you can in fact decide to pull lever *a* and decide to pull lever *b* (but you can't even decide to pull lever *c*). Suppose you are certain that you can decide to pull lever *b* and decide to pull lever *c* (but certain that you *can't* even decide to pull *a*). Finally, suppose that deciding to pull *a* and deciding to pull *c* are great, but deciding to pull *b* is terrible. **Objective-Subjective Decisions** entails that you ought to decide to pull lever *b*, because that is your only option – it is the only decision that you are correctly certain that you can do. This is the wrong verdict. I think it's clear that there is *no* sense in which you ought to decide to pull lever *b*. Deciding to pull lever *b* is terrible, and there is both a distinct decision that you are certain that you can make which is subjectively better and a distinct decision which you can actually make which is subjectively better.

[14] I note that I end up endorsing **Subjective Decisions** in Chapter 3. My point here is that it doesn't satisfy the two constraints on options.

3. Subjective Actions Redux

The constraints look difficult to satisfy in a plausible way; I think something's got to give. In this vein, I recommend dropping the Objective Constraint but retaining OIC. After all, the sole motivation for the Objective Constraint is that it ensures OIC. So there is at least no harm replacing the Objective Constraint with OIC. I think a plausible account satisfying these constraints can be formulated.

In particular, I propose starting with a minimal account of options satisfying the Certainty Constraint. Then it looks easier to satisfy the other constraint (now OIC rather than the Objective Constraint). Recall that the minimal account is as follows:

> **Subjective Actions**. Options are all and only the actions that the agent is certain that she can do. (For some precisification of "actions".)

The challenge is to explain how it satisfies OIC: how does it generate prescriptions that the agent can fulfil? Obviously, an answer to this question involves more than giving an account of options but also looking at how decision theory generates prescriptions from the option set. So far, I've been assuming that decision theory entails that an agent ought to realise the best option – my solution will involve a slightly different formulation of decision theory.

Now, it looks like **Subjective Actions** violates OIC because it seems possible that the subjectively best option for an agent is an action that the agent can't perform. For instance, consider the following case:

> Barry knows he has three levers in front of him *a*, *b* and *c*. He also knows that lever *a* delivers the most money whilst lever *c* delivers the least. Barry is certain that he can pull only one of these levers, and he is certain, for each lever, that he can pull it. However, there is in fact an evil demon that would strike him down just as he is about to pull lever *a*.

It looks like **Subjective Actions** says that Barry's option set contains the pulling of each lever.[15] So given that pulling lever *a* is best, decision theory will say that Barry ought to pull lever *a* – but he can't, hence OIC is violated. Note that this little argument (for the tension between **Subjective Actions** and OIC) assumes that decision theory says that an agent ought to realise the best option. Call this *the naïve formulation of decision theory*. Now it's natural to think that in Barry's case he ought to *try* to pull lever *a*, and indeed, it appears that he can do this – for the case is most naturally understood so that the demon would strike him down just before Barry's trying would be successful. This suggests that we adopt an alternate formulation of decision theory that says an agent ought to *try* to realise the best option. The problem with this is that there's a revenge worry: what happens when Barry can't even try to pull lever *a*? My favoured formulation of decision theory says that an agent ought do as much as she can of the best option, where this is read so that Barry ought to try to pull lever *a* in the case above. Call this *the sophisticated formulation of decision theory*.[16]

To be a bit more precise about the sophisticated formulation, it says, where an agent's best option is A, that she ought to do all B such that: B is a part of A and the agent can B. (I am assuming a simple model of the metaphysics of actions, according to which, deciding to A is part of trying to A, and trying to A is part of A-ing, where A is some outer action.) For instance, suppose an agent's best option is *ford the creek*. The sophisticated formulation says that the agent ought to do any action that is part of fording the creek and such that she can do it. If the agent can only decide to ford the creek (perhaps she would get washed away by the creek before fording it), then the sophisticated formulation says that the agent ought to decide to ford the creek. If the agent can't even decide to ford the creek, then the sophisticated formulation says that there is *nothing* that the agent ought to do (not that she ought to do nothing, but that there simply isn't anything that she ought to do).

[15] Here I assume a precisification of "actions" in **Subjective Actions** as "outer actions". Nothing hinges on this.

[16] For now I'll ignore cases where options are tied. See fn.19 for treatment of ties.

The combination of **Subjective Actions** and the sophisticated formulation satisfies both constraints. The question is whether it delivers sensible verdicts in concrete cases. Recall Barry's case. Barry knows he has three levers in front of him – *a*, *b,* and *c*. He also knows that lever *a* delivers the most money whilst lever *c* delivers the least. Barry is certain that he can pull lever *a*, *b*, and *c*. However, there is in fact an evil demon that would strike him down just as he is about to pull lever *a,* where this is understood as striking him down just before his trying to pull lever *a* would be successful. It looks like **Subjective Actions** says that Barry's options are {pull lever *a*, pull lever *b*, pull lever *c*}.[17] Then the sophisticated formulation entails that Barry ought to try to pull lever *a*. He ought to try to pull lever *a* because he ought to do as much as he can of pulling lever *a*, which amounts to trying to pull lever *a*. If Barry's case is changed so that Barry can pull lever *a*, then the sophisticated formulation entails that Barry ought to pull lever *a*, because doing as much as he can of pulling lever *a* amounts to pulling lever *a* in this case. So far, so good.

The difficult cases are ones where, for instance, Barry can't even decide to pull lever *a*. In such a case, assuming his options are {pull lever *a*, pull lever *b*, pull lever *c*}, the sophisticated formulation entails that there is nothing that Barry ought to do. However, I think that this gets these cases exactly right. For there is nothing that Barry ought to *do*. It's not the case that Barry ought to decide to pull lever *a*, for he can't; it's not the case that he ought to do anything else, because they are all subjectively inferior. So there is nothing that Barry ought to *do*. (It seems that he ought to be about to pull lever *a* and get struck down by the demon – but this isn't something he ought to *do*.)[18] [19]

[17] As before, here I assume a certain precisification of "actions" in **Subjective Actions**. Nothing hinges on this.

[18] This idea about there being nothing that Barry ought to do comes from Hedden (2012), though he uses it for different purposes – see Appendix 1 for more details.

[19] I say the agent ought to do as much as she can of the best option. So far I've been assuming that there is an option that is best. In cases where multiple options are maximal (i.e. tied for bestness), then the sophisticated formulation says that, for each maximal option, it is permissible for the agent to do as much as she can of it. I like this fleshing out of the sophisticated formulation because it means that, like "ought", "permissibility" implies "can".

The Objective-Subjective Puzzle is a puzzle for any account of options. In response, I recommend **Subjective Actions** – options are all and only the actions that the agent is certain that she can do – and the sophisticated formulation of decision theory – according to which, the agent ought to do a much as she can of her best option. This satisfies the Certainty Constraint and OIC, and it delivers sensible verdicts in concrete cases.

4. More on Decisions

In the remaining sections I'll elaborate on some crucial elements of the preceding discussion: on construing options as decisions (section 4), and the objective and subjective constraints at the heart of the puzzle (sections 5 and 6). Finally, in the appendices, I'll compare my discussion to the discussions in Hedden (2012) and Schwarz (2017).

In section 2, I argued that an appeal to decisions won't provide a solution to the puzzle. It was crucial to my argument that the agent might be mistaken about her abilities to make decisions. In this section I want to reject an argument in Hedden (2012) that seeks to establish that the agent is certain and correct about which decisions she can make. More precisely, he argues for the following two claims:

(i) if the agent can decide to ϕ, then she is certain that she can decide to ϕ;

(ii) if the agent can't decide to ϕ, then she is certain that she can't decide to ϕ.

Hedden argues for (i) and (ii) on the basis that whether an agent can make a certain decision depends solely on her mental state. For instance, perhaps an agent can't make a certain decision just when she believes that the decision would be inefficacious. If which decisions an agent can make depends solely on her mental state, then the agent should be certain and correct about which decisions she can make because she is certain and correct about her own mental state. In other words, the agent knows which decisions she can make because, first, she knows her own mental state, and, second, which decisions she can make depends on her mental state (Hedden 2012, pp.352-4).

There are two major problems with Hedden's argument. First, it's clearly false that which decisions an agent can make depends solely on her mental state. For consider an evil demon ready to strike an agent down just as she is about to decide to perform a certain action. The existence of such a demon is not a feature of the agent's mental state and yet whether she can make the decision depends on the existence of the demon. Thus the agent is not guaranteed to know which decisions she can make simply by knowing her mental state.

Second, even if it were true that which decisions an agent can make depends solely on her mental state, then the agent might be uncertain about her mental state and so uncertain about whether she can make the relevant decision. That's because mental states seem vulnerable to Williamson's anti-luminosity arguments (see Williamson 2000, ch.4); also, if the content of mental states is wide, thus implying certain environmental conditions, then the agent may not know which mental states she's in if she's unaware of these environmental conditions.

Hedden is aware of something like these objections to his account and has a response. Before examining his response, it's worth being a bit more precise about Hedden's original argument for why an agent is certain and correct about which decisions she can make. His argument is as follows:

> (P1) Which decisions the agent can make depends solely on her mental state;
>
> (P2) The agent is certain and correct about her own mental state;
>
> (C) Therefore, the agent is certain and correct about which decisions she can make.

I've objected to this argument by objecting to (P1) – I say that there might be an evil demon which interferes with the agent's decision-making. I've also objected to

(P2) – I say that the agent might be uncertain about her own mental state. In response, Hedden (2012) writes:

> 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 ϕ. Plausibly, you lack the ability to decide to ϕ, even if you believe that, were you to decide to ϕ, you would ϕ. 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. (p.354)

He goes on to reply:

> 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 ϕ, 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 the first place. Suppose that the decision to ϕ 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 ϕ. What ought you to do in this case? Certainly, it is not that you ought to make some decision other than the decision to ϕ, since all such decisions look inferior. And it is not the case that you ought to decide to ϕ, 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 ϕ, 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. (p.354)

Here we have a defence of (P1) from my objection. (P1) says that which decisions the agent can make depends solely on her mental state. I objected to (P1) by saying that there might be an evil demon ready to strike the agent down just as she is about to make a decision. Hedden's reply is that these agents, agents whose decision-making is hindered, are irrelevant – the rational 'ought' doesn't apply to them. If we ignore such agents, (P1) holds.

There are a number of problems with this response. First, the rational 'ought' clearly does apply in some cases where an agent's decision-making is hindered. Hedden's example is well-chosen: he considers an agent whose subjectively best decision is hindered. However, suppose that an agent believes that she has a range of decisions open to her, but that (in actual fact) a demon is ready to strike her down just as she is about to make a decision D. Suppose a distinct decision D* looks best and there is no demon ready to strike her down just as she is about to make decision D*. Then the agent ought to make the decision D* – after all, this is what is subjectively best and it is something she can do. So here is a case where the agent's decision-making is hindered but there is something she ought to do.

Nevertheless, let's suppose that we can ignore all agents whose decision-making is hindered. Then there is a second problem for Hedden. For if (P1) holds only for a subset of agents, then the inference from (P1) and (P2) to (C) no longer holds. To see this, suppose that Harry's decision-making is in fact unrestricted, so he is a relevant agent according to Hedden. But suppose that he's unsure whether there is an evil demon that will strike him down just as he is about to make a certain decision. Then regardless of whether he is certain and correct about his mental state, Harry is not certain and correct about which decisions he can make.

There is also a third problem for Hedden. His response is a response to the objection to (P1). It does nothing to address the objection to (P2). (P2) is false because an agent might be uncertain about her own mental state. Excluding agents whose decision-making is in fact unrestricted still leaves some agents who are not certain and correct about their mental state.

I have objected to Hedden's argument for (C) via (P1) and (P2). Is there an alternate argument for (C)? One thing Hedden might say is *that cases where (C) is not true are far-fetched or even impossible on further examination*.

This argument for (C) looks strongest when we consider cases where (unbeknownst to her) the agent's decision-making capacities are restricted by an evil demon. These can appear far-fetched: cases that rarely occur in real-life. Moreover, perhaps such cases aren't even possible. For instance, in an indeterministic world perhaps an evil demon can't predict that you're about to make a certain decision.[20]

However, the problem with this sort of argument for (C) is that there clearly are cases which are possible and not far-fetched, where (C) fails to hold. For consider a case where the agent *believes* that there *might* be a demon that would strike her down just as she is about to make a decision to raise her arm. There need be no such demon. The agent simply assigns some credence to this possibility. This is neither far-fetched nor impossible.

Moreover, the appeal to a supernatural being is dispensable. As mentioned above, there are other more mundane phenomena that can play the role of the evil demon. First, there are pathological psychological phenomena such as phobias, addictions, and indecisiveness. Suppose that the agent is hydrophobic. Then she might be unable to bring herself to genuinely decide to swim in a nearby river. Second, if the agent has a visceral dislike of, say, torturing kittens, then she might be unable to decide to torture kittens. Third, the agent's own beliefs in the inefficacy of her decisions might also play the role of the evil demon. For instance, suppose the agent believes her decision to stop smoking would be inefficacious. Then she might be genuinely unable to decide to stop smoking.[21]

5. More on "Ought" Implies "Can"

The Objective-Subjective puzzle rests on OIC. The motivation for OIC is that it seems bizarre to ascribe an obligation to an agent who cannot fulfil that obligation.

[20] See Graham (2011a) and references contained therein for discussion.

[21] An alternate argument for construing options as decisions would say that (C) is at least satisfied in a wide range of cases, and that this redounds to its credit. See Weirich (1983) for this view. I think it's also implicitly the view in Pollock (1983). This may be true, but this is at best a fall-back position if there is no solution to the Objective-Subjective Puzzle, and as I argue above, *there is* a solution to this puzzle.

Even if I am certain that I can cure cancer, if I can't cure cancer, then it is odd to say that I ought to cure cancer. It may be that I ought to *try* to cure cancer, but it is weird to say that I ought to cure cancer. There have been, however, some criticisms of OIC. This is mostly in the context of the moral *ought* but the same considerations transfer across to the prudential case. In this section I will defend OIC against these criticisms. But first, I'll detail the specific version of OIC I wish to defend.

5.1 Formulation of "Ought" Implies "Can"

Roughly, OIC says that if an agent ought to A, then she can A. However, as it stands, this is ambiguous, because there are omitted time-indices. First, *oughts* are at times. When I promise on Monday to mow your lawn on Tuesday, I ought *on Monday* to mow your lawn on Tuesday. Second, *cans* are at times. When we're talking on your lawn on Monday, I can *on Monday* mow your lawn on Tuesday; but when I take a flight to a faraway country on Monday evening, I can no longer mow your lawn on Tuesday. Which times-indices are relevant to OIC?

As I understand it, decision theory says that an agent acquires an obligation at the time the agent faces a choice. This is a time slightly before she does anything – before she makes a decision. It is roughly the time at which the agent may have asked herself "what ought I do?". Decision theory is silent about what happens to this obligation after she acquires it. Given this, the *ought* in the relevant version of OIC is indexed to the time at which the agent faces a choice. In other words, the relevant version says that if an agent *acquires* an obligation to perform some action, then she can perform that action.

Now all that's left is to index the *can*. I say that the *can* is indexed to the same time as the *ought*. This makes OIC a synchronic principle in that the *ought* and the *can* are indexed to the same time. Of course, the action in question may be an action that is performed at a later time (recall that I'm thinking of actions as coming with a time). So in a sense the principle is diachronic, but it is synchronic in the sense that it relates *ought* and *can* at the same time. This indexing of *can* makes sense given the motivation for OIC: it is bizarre that an agent acquires an obligation-at-a-time-t to do an action that she can't-at-t do.

Given all of this, the formulation of OIC is: (where t_C is the time at which the agent faces a choice)

OIC*. Necessarily, if an agent ought-at-t_C to A, then she can-at-t_C A.

(Henceforth, when I rely on this specific formulation, I will use "OIC*", otherwise I will keep with "OIC".) It's crucial that this version of the principle constrains only an agent's obligation *at the time she faces a choice*, where this is understood as a time before she has done anything. It contrasts to a formulation that constrains obligations *at all times*.[22] This is important for the following alleged objection to OIC.

5.2 Objection 1: Self-Imposed Inability

One alleged problem for OIC is that it entails that an agent who is obliged to perform an action A but then culpably becomes unable to perform A frees herself of the obligation to A – this allegedly makes it too easy to escape one's obligations. For instance, suppose Bill promises on Monday to pay a loan back to you on Thursday. So he ought on Monday to pay back the loan on Thursday. However, suppose he frivolously gambles all his money away on Tuesday so that he can no longer pay back the money on Thursday. It's alleged that OIC entails that it's false that Bill ought on Wednesday to pay you the money on Thursday. This appears to make it too easy to escape one's obligations. This is *the problem of self-imposed inability*. This is what Zimmerman (2008, p146) refers to as the biggest challenge to OIC.[23]

Adapting this objection for the prudential case, suppose Jill ought to two-box in Newcomb's Problem but just before the time for two-boxing comes she stupidly drinks some poison that incapacitates her so that she can no longer two-box. You

[22] For instance, Vranas (2007) and Howard-Snyder (2006) defend formulations that put a constraint on what the agent ought to do *at all times*. Vranas' principle reads: Necessarily, for all times t, if an agent ought-at-t perform A, then she can-at-t perform A.

[23] See Zimmerman (1996), Haji (1997), Howard-Snyder (2006), and Vranas (2007) for discussion. I should note that Zimmerman (1996) thinks this challenge can be met.

might think that Jill doesn't escape her obligation to two-box that easily. Yet given OIC, it appears that once she drinks the poison, then she no longer has that obligation.

However, the version of OIC I need – OIC* – is silent about what becomes of Jill's obligation to two-box when she drinks the poison. Given that Jill ought to two-box, OIC* entails that *at the time she faces a choice*, before she has done anything, Jill can-at-that-time two-box. This says nothing about how Jill's obligations change after this time – and in particular, it says nothing about what happens to her obligation to two box when she drinks the poison. So there is no objection here to the weak sort of OIC that I need.

5.3 Objection 2: Frankfurt Cases

So-called Frankfurt cases have been proposed as counterexamples to OIC.[24] Here is Frankfurt's original example:

> Suppose someone—Black, let us say—wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones's initial preferences and inclinations, then, Black will have his way. . . . (Frankfurt 1969, p.835)

For instance, suppose that Black wants Jones to do something *prima facie* irrational, for instance, spending £5 for a small chance of winning £6 (where "winning £6" is understood as inclusive of the staked £5). As it turns out, Black never has to intervene because "Jones, for reasons of his own, decides to perform and does perform the very action Black wants him to perform" (p.836). It seems that Jones is

[24] For discussion on Frankfurt cases and OIC, see Widerker (1991), Yaffe (1999) (2005), Copp (2008) (1997), Vranas (2007), Haji (1997), Schnall (2001). For discussion specifically in the prudential case, see Wedgwood (2013).

blameworthy for his action even though Black has ensured that Jones could not have acted otherwise. Frankfurt considered such cases as counterexamples to the *Principle of Alternate Possibilities* (i.e. an agent is blameworthy for what she has done only if she could have done otherwise). Additionally, Widerker (1991) argues that it is a counterexample to OIC. For if Jones is blameworthy for spending money, then it seems that he ought not spend the money. So by OIC Jones can not spend the money. But it's false that he can not spend the money, for Black ensures that Jones could not have done otherwise.

The Frankfurt literature is huge and there's little prospect of resolving it here. However, I'll briefly outline my favoured response to show that OIC is at least not obviously-wrong-because-of-Frankfurt-cases.

OIC says that an agent ought to perform an action A only if she can perform A. The motivation for this is that there is something bizarre about attributing to an agent an obligation to do something that she can't do. Now consider Jones. I agree that Jones has an obligation not to do something. However, following Yaffe (1999) (2005), I say that there is no good reason why an obligation *not* to do something entails an obligation *to do something.* An obligation *not* to do something draws a line which no one should cross; an obligation *to do* something demands an action (Yaffe 2005, p311). In particular, there is no good reason why Jones having an obligation *not* to accept the bet entails that he has an obligation *to perform* the action of not accepting the bet. If this is right, then the fact that Jones ought not accept the bet doesn't clearly entail anything about the truth or falsity of OIC, because OIC is a principle about obligations to act.

Indeed, I can say something stronger: it seems that there is good reason to think that having an obligation not to act *does not* entail an obligation to act in a certain way. As Yaffe (1999) (2005) argues, an obligation not to steal does not entail an obligation to perform an action of a certain kind. This is evidenced by the fact that we have an obligation not to steal in our sleep, when we're unconscious, and hence unable to act at all. If the obligation not to steal did entail an obligation to do something else, then we would be violating such an obligation, because we can't act at all in our sleep. But clearly, we are not violating any such obligation when we sleep. So there is good reason to think Jones' obligation not to act in a certain way *does not* entail an obligation to act in a certain way.

However, the case of Jones may provide an objection to OIC in a less direct way – in a way other than constituting a counterexample to that principle. For it's natural to think that similar motivations justify both OIC and the following analogue principle about obligations not to act:

> (*) If an agent ought not A, then the agent can not A.

If the same motivation justifies both OIC and (*), and (*) is counterexampled by the case of Jones, then this would gives us pause about OIC.

However, if (*) is counterexampled by Frankfurt cases, then it *can't* share the same motivation as OIC. For I motivate OIC by appeal to the fact that it is bizarre to attribute an obligation to an agent to perform an action A if she can't A. So if (*) is counterexampled by Frankfurt cases, where it is perfectly natural to attribute an obligation not to act in a certain way, then it *does not* share the same motivation as OIC. So Frankfurt cases do not provide an objection to OIC in this less direct way.

Thus far I've been discussing Frankfurt cases as counterexamples to OIC. But there is another way in which they are relevant. They are useful test cases for a theory of the 'ought'. So what does my conception of decision theory entail in Jones' case? On a natural way of fleshing out of Jones' case, he is certain that he can pay £5 for a small chance of winning £6 and certain that he can refrain from doing this (and there is no other outer action that he is certain that he can do). So it appears that **Subjective Actions** says that his options are *pay* and *refrain from paying*.[25] Given the sophisticated formulation, decision theory will then entail that he ought to do as much as he can of refraining to pay. However, he can't even decide to refrain from paying – because Frankfurt will swoop in and make sure he decides to pay. So I'm

[25] Here I assume a certain precisification of "actions" in **Subjective Actions** which would make these outer actions options. Nothing hinges on this..

committed to saying that there is nothing that he ought to do. This seems right to me, for reasons already explained, namely, that it seems that Jones ought to be such that he would have refused the bet had Black not intervened, but this is not something he ought *to do*.

You might think that there is a problem here for my conception of decision theory, because the intuition is that Jones ought not spend the money, and this is something that isn't entailed on that conception. However, this is because I've been concerned only with deriving obligations to act rather than obligations not to act. I've said nothing so far that commits me to a view about whether there is an obligation not to do something in Jones' case. As a provisional thesis, I say that an agent ought not do any option that is sub-maximal (i.e. ought not do any option that receives a worse EU than another option). This entails that Jones ought not spend the money. (Henceforth, I will understand the sophisticated formulation of decision theory as entailing that an agent ought not perform any option that is sub-maximal. The sophisticated formulation now says what an agent is obliged to do, what she is permitted to do, and what she ought not do. See fn.19 for permissibility.)

5.4 Objection 3: Dilemmas

You might object to OIC on the basis of the existence of dilemmas.[26] The idea is that in a dilemma the agent has both an obligation to A and an obligation to B, where A and B aren't jointly performable. Assuming an *agglomeration principle*, the agent ought to A-and-B. However, the agent can't do this. Hence a counterexample to OIC. A typical example (from Sartre 1980) involves a young man who is deliberating about whether to stay at home to take care of his ailing mother or to leave home to fight with the resistance against an unjust regime. It is natural to think that he both ought to fight with the resistance and that he ought to stay home.[27]

[26] See Vranas (2007, pp.189-190), Nair (2015) and references contained therein.

[27] See Vallentyne (1989) for more stock examples.

In response, I say that there are serious questions marks over whether there are dilemmas in the decision theoretic framework. Considering Sartre's example, it seems to me that in the decision theoretic framework the young man is caught between equally good options and so both are permissible. This is reflected in the theory itself (or at least a naïve application of it): the young man ought to realise a maximal option; if there are two equally good options, then both are permissible.

Nevertheless, let's suppose that dilemmas are to be incorporated into a completed decision theory (and that this can be done without losing the spirit of decision theory).[28] Even in a decision theory with dilemmas, there would still be something like an Objective-Subjective puzzle. For if options are actions that the agent is certain she can do, then even if OIC is not required, it's not plausible that decision theory's prescription would be simply whatever option (so conceived) has greatest EU. This would open the door to lots of crazy *oughts*. For instance, if the agent is late for work and believes she can fly at supersonic speed to work, then decision theory will recommend that she does that. It's hard to believe that the prescription of decision theory is so unconnected to the agent's actual abilities as this. There would be *some* constraint linking what the agent ought to do with what she can do. This constraint wouldn't be something as strong as OIC, as the existence of dilemmas teaches us. However, dilemmas arise in quite special circumstances, where for instance, all the options are equally good. So there would nevertheless be the following constraint: OIC-*in-circumstances-that-do-not-produce-a-dilemma*. This can play the role that OIC does in the Objective-Subjective Puzzle. So even if OIC doesn't hold, there is still a puzzle here.[29]

[28] See Slote (1985) and Norcross (1995) for discussion of incorporating dilemmas into utilitarianism without losing the latter's spirit.

[29] Another common objection to OIC is that pathological psychology proves a counterexample to OIC. For instance, suppose I have a phobia of water. I ought to jump in and save the drowning baby but I can't (because of my phobia). See Graham (2011b) for an argument along these lines. However, see Vranas (2007) and Wedgwood (2013, p.76) for persuasive arguments against this objection.

6. More on the Certainty Constraint

The Certainty Constraint, like OIC, has come in for some criticism. So here I'll defend it. (Given that **Subjective Actions** is a minimal account that satisfies the Certainty Constraint, another way of reading these criticisms is as criticisms of **Subjective Actions**, but I'll frame the criticisms as criticisms of the Certainty Constraint.)

Both Hedden (2015) and Schwarz (2017) object to the Certainty Constraint on the basis that it is *rarely satisfied*.[30] There are two ways to flesh this out. Let's go through each way in turn.

First, you might worry that the Certainty Constraint is rarely satisfied because it requires *certainty* about what is a contingent matter viz. the agent's ability to perform an action. In response, I say that the motivation I gave for the Certainty Constraint (namely, to avoid misevaluating options) motivates something slightly weaker than the Certainty Constraint. I offered the Certainty Constraint to simplify matters and because the difference between it and the weaker constraint didn't matter until now. This weaker constraint avoids the worry about the Certainty Constraint.

The weaker constraint says that the agent is certain of being able to do each of her options, *ignoring circumstances which prevent her from doing all of the options*. For instance, if an agent assigns some credence to having a heart attack in the next few seconds, then she is not certain she can realise each of the alleged options; but this is a doubt about being able to do any of the alleged options, so it does not rule out any of the alleged options according to the weaker constraint. The reason that this weaker constraint is the correct constraint is that the motivation for the Certainty Constraint was to prevent decision theory misevaluating options by ignoring the agent's doubts about being able to perform the option. But decision theory's evaluation is a *ranking* of options – it is a *comparative* evaluation. So if decision theory's

[30] Even though Hedden (2012) accepts it, Hedden later rejects it, see his (2015, fn.30).

evaluation of each option ignores the agent's doubts about being able to do each of them in the same way, then there will be no misranking – all options will be "misevaluated" in the same way.

The worry was that the Certainty Constraint is rarely satisfied because the constraint requires *certainty*. However, the weaker constraint allows that the agent is uncertain of being able to do each option. All it requires is that the agent is symmetrically sceptical of her ability to do each option. So the weaker constraint escapes the scepticism towards the Certainty Constraint – and it is, strictly-speaking, this weaker constraint that is part of the puzzle. For simplicity, however, I will talk as if the Certainty Constraint rather than this weaker constraint is part of the puzzle.

There is a second way of fleshing out the objection that the Certainty Constraint is rarely satisfied. It starts with the following case:

> One day, Brenda hikes from her home and into the countryside. After a while, she comes to a raging creek which blocks her path. Brenda would like to get to the other side of the creek so that she can continue her hike. Brenda is 50-50 on whether she is able to decide to ford the creek. That's because she thinks that Black might strike her down just as she is about to decide to ford the creek, thus preventing her from deciding to ford the creek. In contrast, Brenda is certain that she can go home.[31]

What are Brenda's options? It looks like go home is one option. It also looks like there should be a ford-the-creek-like option. But what is it? It doesn't look like anything ford-the-creek-like satisfies the Certainty Constraint. After all, the agent is not even certain that she can decide to ford the creek. If there's no ford-the-creek-like option, then decision theory would trivially say that Brenda ought to go home, but this doesn't seem right (we might, for instance, flesh out the case so that Brenda really wants to get to the other side of the creek, so that the intuition is that she ought to do something ford-the-creek-like).

[31] This is a development of a case in Hedden (2012) which in turn is a development of a Frankfurt case.

I think this is a problem for the Certainty Constraint. However, I think it is a problem for *everyone*. For it's difficult to say what the ford-the-creek-like option is, *regardless of any alleged constraint on options*. It can't be *decide to ford the creek* or *ford the creek* – these are misevaluated because a decision theoretic evaluation ignores Brenda's doubts about being able to do them. But what else is there? It seems like there isn't anything else.

Given this, I don't think it's motivated at this stage to reject the Certainty Constraint on the basis of Brenda-type cases. What I'm going to do is put aside Brenda-type cases for now. I will explore these cases again in Chapter 4. Let me be clear about exactly what sort of cases I'm putting aside here. The crucial feature of Brenda-type cases is that the agent is uncertain that she can make a decision to perform some action A, where by "uncertain" I mean "assigns a credence between 0 and 1 exclusive". This is the crucial feature, because: if Brenda is certain that she can't decide to ford the creek, then it's unobvious that there is a ford-the-creek-like option (in fact, I would say that it's *obvious* that there *isn't* a ford-the-creek-like option). Moreover, if Brenda is certain that she *can* decide to ford the creek, then the Certainty Constraint doesn't rule out *decide to ford the creek* as an option. So it's essential to Brenda-type cases (as problem cases for the Certainty Constraint) that the agent is uncertain (credence between 0 and 1 exclusive) that she can make a decision to perform some action A. So in putting aside Brenda-type cases, I am assuming that, for every decision, the agent is either certain she can make it or certain that she can't make it.

It is important to get clear on the sort of cases I'm putting aside because you might think that my objection, in section 2, to using decisions to solve the Objective-Subjective Puzzle relied on such cases. So if I'm putting these cases aside, you might think that I no longer have an objection to the decisions proposal. However, my objection to such a proposal *did not* rely on Brenda-type cases. That's because my objection relied on the fact that the agent might be *mistaken* about which decisions she can make. For instance, I objected to **Objective Decisions** (options are all and only the decisions that the agent can make) on the basis that an agent might not be certain that she can make a decision that she can in fact make, so the account doesn't satisfy the Certainty Constraint. Even ignoring agents like Brenda, there are still agents who are mistaken about which decisions they can make. For instance, an agent who is *certain* that she can't make a decision that she can in fact make. I'm not putting aside this sort of case, because, recall, it is essential to Brenda-type cases that the agent is *uncertain* that she can make a certain decision (in the sense that she assigns *nonextreme* credence to being able to do so). So even putting aside Brenda-type cases, **Objective Decisions** doesn't solve the Objective-Subjective Puzzle. Similar remarks go for the other decisions-focused proposals: **Subjective Decisions** and **Objective-Subjective Decisions**.

7. Conclusion

In this chapter I've addressed a puzzle for accounts of options. One the one hand, it appears that options must be actions that the agent can do, so that the prescription of decision theory satisfies 'ought' implies 'can' (OIC). On the other hand, it appears that options must be actions that the agent is certain that she can do, so that decision theory evaluates the options correctly. These constraints look difficult to satisfy whilst simultaneously delivering sensible prescriptions in concrete cases. In effect, this is a tension between two ways of looking at an agent's options: as determined from the *inside* or as determined from the *outside*.

I recommend dropping the Objective Constraint (that an option must be such that the agent can do it) and retaining OIC. This opens up room for the following solution. Options are all and only the actions that the agent is certain she can do (for some precisification of "action"). This is **Subjective Actions**. This trivially satisfies the requirement that the agent is certain she can perform each option. However, it looks like it violates OIC. In response, I propose a sophisticated formulation of decision theory. Decision theory says not "the agent ought to realise the best option" but "the agent ought to do as much she can of the best option". In cases where the agent's beliefs about her abilities are so out of touch with reality that she can't even decide on her best option, then there is nothing that the agent ought to do. This seems exactly right to me.

So in this chapter, I've defended a subjective account of options: they are determined from the agent's perspective. In the next chapter, I look at how to refine "actions" in **Subjective Actions**. My conclusion will be that "actions" should be understood as "decisions". So although a conception of decisions doesn't solve the Objective-Subjective Puzzle, there are reasons to conceive of them as decisions – the reasons, however, are of a very different kind to those operating in the Objective-Subjective Puzzle. Then in Chapter 4 I will pick up a loose end from this chapter: Brenda-type cases. Before that, I have two appendices below, discussing variants of the Objective-Subjective Puzzle in Hedden (2012) and Schwarz (2017).

Appendix 1 – Hedden (2012)

In this appendix and the next, I look at Hedden (2012) and Schwarz (2017), which discuss similar puzzles to the Objective-Subjective Puzzle. Hedden proposes a third constraint in addition to the Certainty Constraint and Objective Constraint, which I resist. Schwarz proposes replacing the Certainty Constraint with a distinct subjective constraint called Anscombe's Condition, which I also resist. Additionally, Schwarz proposes adding a further constraint– and, in my opinion, genuine constraint – to the puzzle: the Fine-Grainedness Constraint. This looks like it's in tension with the subjective constraint (whether this be Anscombe's Condition or the Certainty Constraint). Moreover, my solution to the Objective-Subjective Puzzle appears to violate the Fine-Grainedness Constraint. So this looks like a problem for me. I suggest three responses to this. One of these responses involves co-opting a part of Schwarz's account of options.

First, let's consider Hedden's puzzle. Hedden's puzzle is that it looks difficult to formulate an account of options that satisfies the Certainty Constraint, the Objective Constraint, *and* the following additional constraint:

> Supervenience Constraint. If A is an option for an agent S, then A is an option for any agent in the same mental state as S.

Hedden's motivation for the Supervenience Constraint is that it ensures – given that decision theory's evaluation of an option supervenes on the agent's mental state – that decision theory's *prescription* supervenes on the agent's mental state. In turn, Hedden's motivation for the supervenience of the *prescription* is twofold. First, he says that this is a version of Internalism about practical rationality, which he endorses (2012, fn.4). He doesn't defend such an Internalism but qualifies his project as one of developing an account of options for the internalist. Second, he argues that because decision theory plays a predictive/explanatory role, we have good reason to endorse Supervenience. Let's consider each motivation in turn.

Consider Hedden's first motivation for Supervenience. Now I'm *not* assuming internalism. One of my interests in exploring the question of options is to determine whether an account of options should supervene on subjective or objective features of the agent's decision problem. So Hedden and I are starting from different places. However, there is some reason, even from Hedden's perspective, to drop the Supervenience Constraint. One of my conclusions of this chapter – that options are all and only the actions that the agent is certain she can perform – is obviously sympathetic to Internalism. This means there is no gain to adding the Supervenience Constraint to the Objective-Subjective Puzzle. For an internalist can consider the Objective-Subjective Puzzle as an argument *for* Internalism rather than smuggling in internalist assumptions into the framing of the puzzle.[32]

There is a wrinkle here. For my solution to the Objective-Subjective Puzzle – strictly-speaking – is not internalist, because it doesn't make the *prescription* of decision theory supervene on the agent's mental state. It doesn't do this because the prescription is determined by two factors: first, the options, and, second, how much

[32] My conclusion is *sympathetic* but not an endorsement of Internalism because it would be naïve, at this stage, to think that further considerations will leave my solution untouched, and I've done nothing to rule out objective elements to these changes. Nevertheless, the account is obviously sympathetic to Internalism.

of the best option the agent can do. For instance, if Harry's best option is A, and Barry's best option is B, but while Harry can A, Barry can only decide to B, then I say that Harry ought to A and Barry ought to decide to B. So my solution doesn't make the *prescription* of decision theory supervene on the agent's mental state. However, I take it that this is a harmless dependence on the environment that even the internalist would accept. The options and ranking of options are determined by the agent's mental state; moreover, the sort of thing the agent ought to do is determined by her mental state – it is just whether the agent ought to decide, try, or actually realise A that is determined by her abilities.

Now let's consider Hedden's second argument. Hedden assumes that decision theory plays a predictive/explanatory role. That is, we use decision theory to predict what an agent will do. (Also we use decision theory to explain why an agent acted as she did – but the argument focuses on the predictive side of the role.)

Hedden says that if you didn't have supervenience of the prescription on the agent's mental state, then you would have a case where two agents in the exact same mental state (but in different external circumstances) are predicted to do two different actions. Hedden illustrates this with Jane and her mental duplicate Twin-Jane: "…we should predict that Jane will ford the creek, while Twin Jane will immediately do an about-face and head straight home" (p.349). About this situation, Hedden says: "But this is bizarre! They are in exactly the same mental state, after all! If they displayed such radically different behaviour, it would appear that at least one of them wasn't fully in control of her actions (and hence not fully rational)" (p.349). As Hedden earlier says, "fully rational agents successfully fulfil 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" (p.344).

So the idea is that if the Supervenience Constraint is false, then two agents in the same mental state will be predicted to perform different actions. If they do these things, then at least one of them isn't in control of her actions. So at least one of them isn't rational in a fuller sense which includes abiding not only by decision theoretic norms but also epistemic norms and being in control of your actions. So the alleged absurd consequence (if the Supervenience Constraint is false) is that in order to be decision theoretically rational an agent would have to forfeit rationality in the fuller sense – in particular, she would have to forfeit rationality in the fuller sense by losing control of her actions.

I think this argument isn't convincing for three reasons. First, it's not clear that the alleged absurd consequence is absurd. The consequence is that for an agent to be decision theoretically rational, she would lose out on some other form of rationality. Perhaps there are agents who are unlucky in this sense. Second, it's not clear that if mental duplicates do different things, then at least one of them isn't in control of her action. Different agents might have a different wiring between their mental states and actions, so, for different agents, the same mental state determines different actions. This seems compatible with both agents being in control of their actions. They might, for instance, repeatedly perform the same action when they're in the same mental state. That seems like enough to say that these agents are in control of their actions. Third, it's not clear that rationality in the fuller sense requires being in control of your action. If rationality in the fuller sense is about having the right *pattern* of mental states and actions, rather than tokening the right sort of *processes*, then there doesn't seem much appeal to control as a necessary condition for rationality in the fuller sense. For an agent is in control of her action in virtue of the right sort of *process* leading up to it. So if rationality in the fuller sense is about patterning rather than process, it doesn't like there's room for control. So for these three reasons, I don't think Hedden's argument is convincing.

In summary, Hedden thinks an additional constraint should be added to the Objective-Subjective Puzzle, namely, the Supervenience Constraint. He offers two motivations. I have resisted both.[33]

[33] Hedden's solution to his puzzle is that options are all and only the decisions that the agent can in fact make. I've already discussed in sections 2 and 4 why I don't think this works as a solution to my puzzle (and, of course, this entails that it won't work as a solution to Hedden's puzzle; that's because Hedden's puzzle contains an additional constraint on options, so it is *harder* to solve than my puzzle).

Appendix 2 – Schwarz (2017)

Anscombe's Condition

Schwarz's puzzle consists of three constraints: the Objective Constraint, Anscombe's Condition, which is the replacement for the Certainty Constraint, and the Fine-Grainedness Constraint, which I'll come on to later. Anscombe's Condition is as follows:

> Anscombe's Condition: If A is an option, then the agent can become rationally certain of A just by resolving her decision problem.

Schwarz attributes this condition to Jeffrey (1968). It's called *Anscombe's* Condition because Anscombe (1963) said that intentional agency provides "knowledge without observation". For instance, when I choose to raise my hand, it's not the case that I know I've raised my hand because I observe it in the same way as I observe other people's hands going up. I know that I've raised my hand because I chose to do so. It's this sort of learning experience that is appealed to in Anscombe's Condition.

Anscombe's Condition is motivated in the same way as I motivate the Certainty Constraint: it is needed to ensure options are evaluated correctly. Schwarz gives the following example. Suppose that Rob knows he can cook risotto and haggis for Alice. He wants to cook Alice's least favourite meal, but he doesn't know what Alice's least favourite meal is. He is fairly confident (but not certain) that her least favourite meal is haggis, but her least favourite dish is in fact risotto. Schwarz says:

> If cooking Alice's least favourite food were an option for Rob, then this is what he ought to choose; it would maximise expected utility. But if Rob falsely suspects that Alice's least favourite food is, say, Haggis, then it may well be rational for him to cook Haggis and thus *not* to cook Alice's least favourite dish (which is risotto). So cooking Alice's least favourite dish is not an option for Rob. (p.5)

Schwarz is saying that *cook Alice's least favourite meal* is not an option for Rob, because if it were, then that is what he ought to do; however, intuitively, Rob ought to do something else, namely, cook haggis. In short, cooking Alice's least favourite meal is not an option for Rob because otherwise decision theory delivers the wrong verdict.

What is it about *cook Alice's least favourite meal* that means it's not an option for Rob? The action appears to satisfy the Objective Constraint. After all, Rob can cook risotto, which is in fact Alice's least favourite meal, so he can cook Alice's least favourite meal. You might think that a stronger sort of 'can' would entail that Rob can't cook Alice's least favourite meal. For instance, you might think that the conditional analysis of "can" will allow us to say that Rob can't cook Alice's least favourite dish.[34] After all, if Rob were to decide to cook Alice's least favourite meal, then presumably (because he is mistakenly confident that her least favourite dish is haggis) he will cook haggis, but this is not Alice's least favourite meal. So on the conditional analysis, Rob can't cook Alice's least favourite dish. Schwarz considers this but rejects it. That's because we can tweak the case so that Rob can cook Alice's least favourite meal, even on the conditional analysis of 'can'. Recall that Rob is mistakenly confident (but not certain) that Alice's least favourite dish is haggis. Tweak the case so that Rob's suspicions are correct, so that Alice's least favourite dish is in fact haggis – but crucially, keep fixed that Rob isn't certain that Alice's least favourite dish is haggis. Then presumably it's true that if Rob were to decide to make Alice's least favourite meal, then he would make Alice's least favourite meal (because if he decided this, then he would make haggis). So on the conditional analysis, Rob can cook Alice's least favourite meal. However, *make Alice's least favourite dish* still shouldn't count as an option.

Schwarz's idea is that *cook Alice's least favourite meal* is not an option for Rob because Rob *couldn't become rationally certain that he'd cook Alice's least favourite meal just by resolving his decision problem*. Whether he decides to cook risotto, haggis, or even Alice's least favourite meal (these decisions would constitute a resolution of the decision problem), it seems that Rob won't be certain that he will cook Alice's least favourite meal. Hence Schwarz endorses:

[34] By the conditional analysis of "can" I mean the account that says an agent can perform some action A just when: if she were to decide on A, then she would A.

> Anscombe's Condition: If A is an option, then the agent can become rationally certain of A just by resolving her decision problem.

Schwarz rejects the Certainty Constraint, or rather, he thinks it is *redundant*, for Anscombe's Condition can do all the work of the Certainty Constraint but not vice versa. Recall the sort of case that motivates the Certainty Constraint: suppose that you're stood in front of two levers, *a* and *b*, that, when pulled, deliver $100. You're given the choice of pulling one of the levers. But here's the catch: sometimes lever *b* doesn't work – it sometimes jams, in which case you wouldn't be able to pull it. You assign 0.5 credence to lever *b* jamming. *Pull lever b* isn't an option here because it would be misevaluated by decision theory's purely consequentialist evaluation. The Certainty Constraint correctly rules out *pull lever b* as an option. This is the motivation for the Certainty Constraint. Schwarz says that Anscombe's Condition can do the work of the Certainty Constraint. For in this case, you would not become certain of pulling lever *b* after you make a decision to pull lever *b* (nor any other decision), so *pull lever b* wouldn't count as an option given Anscombe's Condition.

Moreover, if we flesh out the Rob case so that Rob is certain that he can cook risotto, certain that he can cook haggis, and certain that one of these is Alice's least favourite meal (though he isn't sure which), then it will come out that Rob is certain that he can cook Alice's least favourite meal. So the Certainty Constraint alone won't rule out *cook Alice's least favourite meal* as an option for Rob. So the Certainty Constraint can't do the work of Anscombe's Condition but the latter can do the work of the former. This, Schwarz thinks, suggests that the Certainty Constraint is redundant.

In response, I say that the Certainty Constraint can do all of the work of Anscombe's Condition but not vice versa. Consider the Rob case again (I've introduced a few different versions of this, but it doesn't matter which one is considered). The Certainty Constraint *does* rule out *cook Alice's least favourite meal*, if the 'can' is understood along the lines of the conditional analysis. Rob is not certain that if he were to decide to cook Alice's least favourite meal, then he would cook Alice's least favourite meal. For if he decided to cook Alice's least favourite meal then he would cook one of risotto or haggis, and he isn't sure of either that it is Alice's least favourite meal.

Moreover, consider a case where the agent is deliberating about which of three levers to pull – *a*, *b,* and *c*. Suppose she is certain she can pull *a* and *b*, but certain she can't even decide to pull *c*. Nevertheless, she can decide to pull lever *c*. The Certainty Constraint correctly rules out *decide to pull lever c*, on the basis that it would be misevaluated – for its decision theoretic evaluation would ignore the agent's doubts about being able to decide to pull *c*. However, it seems to me that Anscombe's Condition will not – for the agent could become rationally certain of making the decision to pull lever *c* if she were to decide to pull lever *c*. So I think things are in fact the other way around: the Certainty Constraint can do all the work of Anscombe's Condition, but not vice versa.

So I don't think that Anscombe's Condition should replace the Certainty Constraint in the Objective-Subjective Puzzle. The Certainty Constraint is the genuine subjective constraint.

The Fine-Grainedness Constraint

Schwarz also proposes a third additional constraint. He considers an experienced marksman aiming at a distant target. The marksman takes into account the wind conditions, the quality of the rifle, and other such factors to take the shot in the very specific way that he does. The marksman has fine-grained control over the precise torque in his joints. The marksman's options need to be as fine-grained as these configurations of torque in his joints. For instance, perhaps his options are the torque configurations themselves (or decisions for such things). His options need to be fine-grained because: imagine that the options are coarse-grained, for instance, *aim at the target* and *don't aim at the target.* Then obviously decision theory will say that the marksman ought to aim at the target. But this prescription is too coarse-grained. The marksman would aim at the target with a wide range of torque configurations, but he ought to realise the torque configuration that he thinks would give him the best shot at hitting the target.

This is interesting because (as Schwarz says) it looks difficult to generate fine-grained options whilst simultaneously satisfying a subjective constraint – whether that be Anscombe's Condition or the Certainty Constraint. The prima facie option set is the set of torque configurations. However, if the option set is to satisfy a subjective constraint, then the agent is required to have beliefs about each option. For instance, if the Certainty Constraint is the subjective constraint, then the agent is required to be certain of each option that she can do it. So if the option set is the set of torque configurations, then the agent needs beliefs with very sophisticated contents – contents about specific torque configurations – which looks beyond the agent's cognitive abilities in a typical case.

Here is another case which demonstrates the same thing, this time from Jeffrey (1983, §11.9):

> Example 10: The comforter
>
> The agent is trying to comfort a lady whose cat has been killed. This may consist in any of a variety of acts, such as giving her another cat, holding her hand, or saying "He was getting old and stiff, anyway." And there are many ways of performing the last-mentioned act, of saying the words, some of which would be more likely to produce comfort than others: variations in volume of voice, proximity of speaker to hearer, and facial expression might all be important. The agent might have an accurate sense of how he is speaking the comforting words in each of these respects without being able to verbalize it. Thus, he might be tacitly aware that saying the words in a loud voice from across a room with his facial muscles relaxed might have a disturbing effect; and he might be able to control his distance, volume, and facial expression in the relevant ways; and he might nevertheless be unable to produce or recognize a true description of what he is doing, in terms of distance, volume, and facial expression.

As before, the prima facie options here are fine-grained – they are variations in the way the agent says "He was getting old and stiff, anyway". If the options satisfy a subjective constraint, then it appears the agent has very sophisticated belief contents.

So Schwarz's additional constraint is as follows:

> The Fine-Grainedness Constraint. The options need to be fine-grained in cases like the Marksman Case and the Comforter Case.

This is thought to clash with a subjective constraint because together they demand cognitive abilities beyond the reach of a typical agent.

The problem as it appears for me is that there is a tension between the Fine-Grainedness Constraint and the Certainty Constraint. Unsurprisingly, my account of options – **Subjective Actions**, which trivially satisfies the Certainty Constraint – appears to violate the Fine-Grainedness Constraint. For consider what **Subjective Actions** entails in the Comforter Case. It entails that the agent doesn't have as options specific ways of saying "He was getting old and stiff, anyway". That's because the agent will not be certain that she can realise specific ways of saying "He was getting old and stiff, anyway" – and this, in turn, is because the agent doesn't have beliefs with such sophisticated contents. How can I respond to this?

The first thing to say is that it's not obvious that there really is a problem here. First, one might insist that (in any interesting case) the agent *is* capable of having beliefs with the required sophisticated contents. One way to flesh this out is to say that it doesn't take much to have such beliefs. Second, in (for instance) the Comforter Case, you might find some fine-grained options that can play the role of specific ways of talking and yet do not require sophisticated belief contents. For instance, perhaps one of the options is *speak in such a way that has a very good (objective) chance of comforting the lady*. Another option might be *speak in such a way that has a decent (objective) chance of comforting the lady*, and so on. This would require the agent to have beliefs about chances but this is perhaps less bad than beliefs about torque configurations and the like.

Nevertheless, *if* there is a problem for my account here, I think there is an adequate response. Fortunately, this is already provided by Schwarz. However, it is wrapped up with his account of options. So first, I'll outline Schwarz's account of options. Then I'll show how I can co-opt part of this account to provide a solution to the problem.

Schwarz's account of options is designed to satisfy three constraints: Anscombe's Condition, the Objective Constraint, and the Fine-Grainedness Constraint. First, consider what I'll call the *narrow physiological states* of the agent. These are internal states of the agent. They are typically decisions, but they can also include motor commands to muscles which issue in distinctive bodily stances. These are *narrow* physiological states because their occurrence depends on little in the external environment. Schwarz says that the agent's options *are* the narrow physiological states that she can realise. This proposal trivially satisfies the Objective Constraint. It also delivers fine-grained enough options in the Marksman Case and the Comforter Case. For instance, in the Marksman Case, the options would be specific motor commands that issue in particular bodily stances. So the proposal satisfies the Fine-Grainedness Constraint. The challenge for Schwarz is to explain how it satisfies Anscombe's Condition, which recall is:

> Anscombe's Condition. If A is an option, then the agent can become rationally certain of A just by resolving her decision problem.

Schwarz says that when the agent realises a narrow physiological state (this constituting a resolution of her decision problem) she becomes certain of a specific proposition. This proposition does not necessarily refer to the state but corresponds to it in some looser sense, which I'll explain in a moment. The upshot is that a narrow physiological state (that the agent can realise) satisfies Anscombe's Condition *in a mediated sense* – in the sense that the agent becomes certain of something that *corresponds* to the state. (The mediation does not consist of the agent becoming certain of a proposition; rather, it consists in the agent becoming certain of a proposition *with a content that does not necessarily refer to the narrow physiological state*.)

To give some background, Schwarz thinks that the agent's beliefs should be modelled by a credence function in the normal way, but with one addition. The agent has credences defined over special *virtual propositions*. These propositions are paired with possible narrow physiological states in the sense that, when the agent realises a narrow physiological state P, she becomes certain of the corresponding virtual proposition Q and then adjusts her conditional-credences-given-Q in light of what she observes.

The correspondence (between narrow physiological state and virtual proposition) is not in virtue of the virtual proposition *describing* the physiological state. The virtual propositions may not describe a genuine possibility at all. Instead, the virtual propositions correspond to physiological states by being involved in learning after realisation of a physiological state.

For instance, an agent might observe that after she realises a physiological state P, she generally moves left. So her conditional probability for moving left given the corresponding virtual proposition Q is high. One day there is a wall to her left and realising P doesn't lead to her moving left. So the agent updates her credences such that she assigns a lower credence to moving left given Q. Or, perhaps, if she is sophisticated enough, she will assign a high credence to moving left given the conjunction *Q and there is no wall to my left.*

This picture endows a certain structure to our model of the deliberating agent. The agent has beliefs about the physiological states she can realise, but they are not beliefs that are necessarily about the physiological states *thought of as under a description which picks out the states*. Rather she has beliefs about them under a possibly incorrect description. What makes the belief a belief *about a certain physiological state* rather than some other physiological state is that after realising the physiological state she updates her credences by adjusting her probabilities conditional on the corresponding virtual proposition.

As I said, Schwarz says that options are the agent's narrow physiological states that she can realise. This trivially satisfies the Objective Constraint and delivers fine-grained options. It also satisfies Anscombe's Condition in the sense that the *corresponding virtual propositions* satisfy Anscombe's Condition. When the agent realises a narrow physiological state (and so resolves her decision problem), it's simply a feature of Schwarz's model of the deliberating agent that she becomes certain of the corresponding virtual proposition. So the narrow physiological states satisfy

Anscombe's Condition in a mediated way: the agent becomes certain of a corresponding proposition after resolving her decision problem.[35]

Schwarz's account of options simultaneously satisfies the Fine-Grainedness Constraint and a subjective constraint. I want to co-opt that part of his account responsible for this. The key to Schwarz's account satisfying these two constraints is the posited correspondence between virtual propositions and narrow physiological states. This allows the agent to have beliefs about the narrow physiological states, which are fine-grained, without having sophisticated belief contents – thus undercutting the reason for thinking that there is a tension between the Fine-Grainedness Constraint and a subjective constraint.

In more detail, this is how I wish to alter (or, rather, interpret) my account of options so that it satisfies the Fine-Grainedness Constraint. Consider my account as coming with a claim about our model of the deliberating agent. The agent's narrow physiological states (that she can realise) correspond to virtual propositions in the way Schwarz describes, i.e. in that the agent becomes certain of the virtual proposition upon realising a narrow physiological state. My account of options says that options are all and only the actions that the agent is certain she can realise. I wish to interpret this so that a narrow physiological state that the agent is certain she can realise – in virtue of being certain that she can realise the corresponding virtual proposition – counts as an option on my account. The idea is that these narrow physiological states, just like more mundane actions such as pulling a lever, will count as options. In the Marksman and Comforter cases, this will ensure there are fine-grained enough options.

It's worth seeing the difference between my account (interpreted as just outlined) and Schwarz's. Schwarz's account starts from the outside, as it were. The agent's options are the narrow physiological states that she can realise. He ensures that these satisfy the subjective constraint – Anscombe's Condition in his case – by

[35] The narrow physiological states also have an EU in a mediated way – in the sense that they inherit the EU's of their corresponding virtual propositions.

positing a correspondence between virtual propositions and narrow physiological states. In contrast, I start from the inside: the agent's options are the actions that she is certain she can do. But I want to interpret this such that an agent's options include the narrow physiological states that she is certain she can realise – where this certainty takes a mediated form, namely, of the agent being certain that she can realise the corresponding virtual proposition. The two accounts will come apart in cases where the agent isn't sure, of some narrow physiological state, that she can do it. Schwarz will say that it is an option; I will say that it isn't an option.

To conclude this appendix, Schwarz (2017) proposes an additional constraint on options, which demands that they are fine-grained enough in cases where the agent looks like she doesn't have the conceptual machinery to have beliefs about the actions that are the *prima facie* options. I suggested a couple of easy ways to resolve this (first, maintain that it doesn't take much to have such beliefs, second, find options which require less sophisticated belief contents), but if neither of these proposals work, then I would co-opt part of Schwarz's account of options. That is, I would say that the agent has beliefs about such actions in virtue of her doxastic space having virtual propositions which correspond to the actions – correspond in the sense that the agent becomes certain of the virtual proposition when she realises the corresponding action. (However, to keep the subject matter tractable, in the remainder of the thesis, I will be putting aside cases where the agent doesn't appear to have the cognitive abilities to have beliefs about the *prima facie* options.)

Chapter 3 – What Kind of Action is an Option?

1. Introduction

In Chapter 2 I argued for the following account of options:

> **Subjective Actions**. Options are all and only the actions that the agent is certain that she can do.

This account is schematic in that "actions" awaits interpretation. In this chapter I want to turn to the question of how this should be interpreted. I argue for the following account of options:

> **Subjective Decisions**. Options are all and only the *decisions* that the agent is certain she can make.

As did Chapter 2, this chapter assumes that, for every decision, the agent is either certain she can make it or certain she can't make it. So I end this chapter with an endorsement of **Subjective Decisions** *under that assumption.* (Cases where this assumption fails are discussed further in the next chapter.)

Let me flag how I'm understanding **Subjective Decisions**. As I said in Chapter 1, I'm thinking of actions as being performed at (or over) a particular time. So in particular, I'm thinking of decisions as being made at a particular time. *When* are the relevant decisions? Before answering this, let me give some background. As I said in Chapter 1, I'm thinking of an account of options as generating a set of options for an agent who faces a choice – that is, an agent before she's done anything, and when the question "what ought I do?" has arisen. I've previously labelled this time t_c. I am also assuming that options are *immediately performable.* That is, I assume that they start (and possibly end) immediately after the time the agent faces her choice. I've previously labelled this time t_i. Given this as background, I'm understanding **Subjective Decisions** such that the relevant decisions are decisions that take place *at* t_i. That is, they are minimal actions (i.e. they start and end at t_i).[36] I'll often say that they are made *right now* or *presently* to indicate their temporal immediacy (although, strictly-speaking, they are slightly after the time at which the agent faces the choice). So **Subjective Decisions** says that an agent's options are all and only the decisions-at-t_i that the agent is certain she can make.

For instance, I am certain that I can make a decision tomorrow morning to have cereal for breakfast. That is, I am certain that I can token this decision *tomorrow morning*. This *isn't* an option according to **Subjective Decisions**. That's because the decision takes place tomorrow morning rather than *immediately*. My options according to **Subjective Decisions** are things like: make a decision *immediately* to have cereal for breakfast tomorrow morning; make a decision *immediately* to go for a walk now; make a decision *immediately* to go for a walk in five minutes time etc. I will leave this qualification implicit in what follows.

I argue that rivals to **Subjective Decisions** deliver strange prescriptions in some concrete cases. I consider a *rival* to be an account that conceives of options as some species of outer action.[37] My argument will put aside cases where it's obvious that the options are decisions as per **Subjective Decisions**. For instance, suppose Harry is certain that he can make any decision, but he's unsure that any of these decisions will be effective. Perhaps he thinks an evil demon will strike him down after the formation of a decision but before outer action. Then it's obvious that Harry's options are *decisions*. I put aside these cases and focus on cases where it is less obvious that **Subjective Decisions** is right, namely, where the agent is sure that she can perform at least some outer actions.

[36] Minimal actions were introduced in Chapter 1, section 3. They were contrasted with extended actions, which are actions that takes place over a certain time span.

[37] I'm interpreting "rival" such that the following account doesn't count as a rival: options are all and only the actions (*both* outer actions *and* decisions) that the agent is certain she can do. I don't consider this a rival because it still selects decisions as options. Regardless, this chapter can be read as assuming that this account is off the table for now in order to simplify the discussion to come. What I say in section 4 cuts against this account, as I will make clear. So if one is sympathetic to this account, it will still be ruled out by what I say later.

After this introduction, I begin ruling out rivals to **Subjective Decisions**. First, I distinguish between two readings of "can" – *diachronic* and *synchronic*. I reject any rival that appeals to the diachronic "can". Section 3 is concerned with an argument by Sobel (1983) against three rivals. In short, I appeal to Sobel's argument to rule out further rivals to **Subjective Decisions**. But I understand the significance of Sobel's arguments slightly differently to Sobel himself, so it will take some time to deal with this. Finally, I object to a remaining rival. The conclusion will be **Subjective Decisions**, by virtue of being the last account standing.

In this chapter I'll assume that the agent is correct about what she is certain she can do. Given **Subjective Actions**, i.e. that options are actions the agent is certain she can do, this means the agent can do each of her options. This simplifies the discussion to follow. In particular, it simplifies the formulation of decision theory. In Chapter 2 I recommended the following sophisticated formulation of decision theory: an agent ought to do as much as she can of the best option. This was recommended over a naïve formulation, according to which, an agent ought to realise the best option. Given the assumption that the agent is correct about what she is certain she can do, then the sophisticated formulation reduces to the naïve formulation.

2. Objection to Diachronic Rivals

In this section I reject any rival to **Subjective Decisions** that makes use of the diachronic "can". That is, I reject any account of options that both conceives of options as outer actions and appeals to the diachronic "can". First, I explain the distinction between a synchronic and a diachronic ability (2.1). Then I'll present my objection to diachronic rivals (2.2).

2.1 Synchronic and Diachronic Abilities

For an outer action, and, in particular, for an extended outer action, there is a distinction between being able synchronically to perform it, and being able diachronically to perform it. These two abilities pull apart most dramatically in the case of Professor Procrastinate:

> Professor Procrastinate (henceforth PP) is invited to write a chapter in a prestigious anthology on his area of speciality. He must accept or reject the invitation now. If he were to write the chapter, then he would write the chapter at a later date. Obviously, writing the chapter is really good. However, PP is a procrastinator. If he accepts the invitation, he will in fact not write the chapter but will instead keep putting it off to do more menial tasks. There is nothing standing in the way of PP writing the chapter; it's just that as a matter of fact he won't. Moreover, there's nothing PP can do now to ensure that he doesn't later procrastinate – for instance, PP promising himself to write won't stop him later procrastinating. Now, if PP accepts the invitation but doesn't end up writing the chapter, this would be the very bad, because it would annoy the editors of the anthology. It's much better to reject the invitation than to annoy the editors. (PP knows that all of this is the case.)

Can PP accept-and-write? That is, does PP *now* have the ability to accept-the-invitation-and-write-the-chapter? I take it that there are two intuitions here. The first says that PP can't accept and write because there is nothing PP can do right now to ensure that he writes the chapter. PP can't "commit himself", he can't "bind himself", he can't make certain that he accepts and writes, so he *can't* accept and write. The second says that PP *can* accept and write because it is up to him whether he does so or not. No evil demon will prevent PP from later writing. Whether PP later writes is something that depends only on what he does. If, for instance, PP were suitably motivated both now and in the future, then he would accept and write, so he can accept and write. I consider it obvious that two sorts of ability pull apart in PP's case, and I have done my best to bring out the pull of each. I won't try to analyse these accounts any further, because the rough characterisations of them will do for my purposes here. I will label the "can", on which it is true that PP can't accept and write, the synchronic "can". That's because, on this "can", an agent can perform an action when she can do something right now (i.e. *synchronically*) that would ensure that she performs A.[38] I will label the "can", on which it is true that PP can accept and write, the diachronic "can". That's because, on this "can", an agent can perform an action when she would do it if she were suitably motivated both now and in the future (i.e. if she were suitably motivated *diachronically*).[39]

I will assume that for minimal actions these two abilities do not pull apart. That is, for a minimal action A, an agent can diachronically A just when she can synchronically A. For instance, consider the action of raising your arm right now. If this is something that you can do, then it seems that it is both up to you (i.e. there is nothing stopping you from doing it) and that you can ensure right now that it happens (viz. by raising your arm right now). It's only for extended actions that these two abilities pull apart, because it's only for extended actions that something you can diachronically do (i.e. you would do if suitably motivated both now and in the future) is something that you can't ensure right now, because your future-self might, as it were, get in the way. This is what happens with PP. He can diachronically accept and write but he can't do something right now to ensure he does this because his future-self won't cooperate. In other words, these two abilities pull apart when your future time-slices – who would have to play ball to complete the action – are uncooperative, but for minimal actions, there are no such future time-slices.

So in particular, this distinction between abilities does not matter for S**ubjective Decisions**, because decisions, as I am understanding them, are minimal

[38] As with the decisions appealed to in **Subjective Decisions**, describing the action (which ensures A) as happening *right now* is a simplification – the action happens at t_i. It must be this way because the agent faces her choice before she's done *anything,* and t_i is the time immediately after the agent faces her choice.

[39] See Portmore (2017, ch.3) for more on this distinction between two types of *ability*. I have characterised these two types of abilities in the simple framework of agency that I proposed in Chapter 1.

actions – they are actions that are performed and completed right now. However, it does matter for rivals to **Subjective Decisions**, because these rivals conceive of the options as outer actions, and extended actions are one sort of outer action. So it is natural to wonder whether any one of the abilities is ill-suited to play a role in an account of options, in which case we could rule out all accounts that appeal to that particular "can". Indeed, in the next subsection we will see that we can rule out all rivals that appeal to a diachronic "can".

2.2 Against Diachronic Rivals

In this subsection I will object to any rival to **Subjective Decisions** that appeals to the diachronic "can" – *diachronic rivals*, as I will call them. I will do this by objecting to an account that says options are all and only the outer actions that the agent is certain she can diachronically do – call this **Naïve Diachronic**. I will then extend this to all diachronic rivals. (Recall that a rival makes options outer actions, so a diachronic rival is an account of options that makes options outer actions *and* is one that appeals to the diachronic "can".)

The problem for **Naïve Diachronic** is what it entails in the following two cases (one of which is simply PP's case from before):

> *Professor Procrastinate*. PP is invited to write a chapter in a prestigious anthology on his area of speciality. He must accept or reject the invitation now. If he were to write the chapter, then he would write the chapter at a later date. Obviously, writing the chapter is really good. However, PP is a procrastinator. If he accepts the invitation, he will in fact not write the chapter but will instead keep putting it off to do more menial tasks. There is nothing standing in the way of PP writing the chapter; it's just that as a matter of fact he won't. Moreover, there's nothing PP can do now to ensure that he doesn't later procrastinate – for instance, PP promising himself to write won't stop him later procrastinating. Now, if PP accepts the invitation but doesn't end up writing the chapter, this would be the very bad, because

> it would annoy the editors of the anthology. It's much better to reject the invitation than to annoy the editors. (PP knows that all of this is the case.)
>
> *Professor Uncertain*. This is exactly like PP's case except that Professor Uncertain (henceforth PU) is uncertain that she will live long enough to accept-and-write, because she might die beforehand. PU is sure that she'll live long enough to accept the invitation, and sure that she'll live long enough to reject the invitation. But she's not sure she'll live long enough to write. Crucially, even though she assigns some credence to dying before she gets chance to write, she is certain that *if* she survives until the appropriate time to write, then, just like PP, she won't write, even though there would be nothing stopping her from writing.

Naïve Diachronic entails that the options in PP's case are:

> Accept and write
> Accept and don't write
> Reject and write
> Reject and don't write
> Accept
> Reject

That's because these are all the outer actions that the agent is certain she can diachronically do. Recall that the relevant ability is a diachronic ability, and although PP isn't certain that he can do anything right now to ensure that he performs the extended actions, there is nothing stopping him from performing the extended actions, which is enough for him to count as having the diachronic ability to perform them.

PP thinks that accept-and-write leads to the best outcome (namely, chapter in prestigious anthology), so decision theory would say that PP ought to accept-and-write. Now this verdict may be objectionable, but my point here is not to contest this verdict but to note that it leads to trouble when compared to the verdict in PU's case. For this verdict in PP's case effectively ignores the agent's predicted future irrationality. PP ought to accept and write because that is the best option even though he would irrationally fail to complete this action if he were to embark upon it. Of course, there is a sense in which PP's predicted irrationality is relevant – it will be taken into account in the EU of *accept*, which is also an option. However, there is also a good sense in which it is irrelevant. For the extended action *accept and write* is evaluated without recourse to PP's predicted future irrationality, and that is the option that has the best EU.

Now consider PU. **Naïve Diachronic** entails that the options in PU's case are:

Accept

Reject

That's because these are the only outer actions that the agent is certain she can diachronically do – recall that the agent thinks she might die before the time of writing. Given this, decision theory will say that PU ought to reject. That's because PU thinks *accept* leads to the worst outcome, namely, annoying the editors. So unlike PP, PU's predicted future irrationality is very relevant. She ought to *reject* the invitation because she thinks that if she were to *accept*, then she would end up irrationally procrastinating, and hence end up annoying the editors.

I think this asymmetry is unacceptable. On **Naïve Diachronic**, PP effectively shouldn't take into account his predicted future irrationality but PU should, despite the only difference being that PU assigns some credence to being unable to perform any extended action. What's strange is that PU's doubts about being able to perform extended actions force her options to be minimal outer actions, which in turn makes her predicted irrationality relevant. Of course, her doubts about being able to perform extended actions are relevant, but they shouldn't make her predicted future irrationality relevant *as well* – at least, they shouldn't given that PP's predicted future irrationality is effectively irrelevant. So **Naïve Diachronic** ought to be rejected on the basis that it leads to weird asymmetries in the relevance of the agent's future predicted irrationality.

In contrast, compare **Naïve Diachronic** with the analogue synchronic account, which says that options are all and only the outer actions that the agent is certain she can *synchronically* do. This account says that PP's and PU's option set is {accept, reject}. This is because *accept* and *reject* are the only actions that each agent is certain they can synchronically do. For instance, PP is not certain he can synchronically accept-and-write, because there's nothing he can do now to ensure that he writes, so that doesn't count as an option. This means that both PP and PU ought to reject the invitation, which in each case makes their predicted future irrationality relevant. There is no weird asymmetry here.

So I think **Naïve Diachronic** is to be rejected. Moreover, any diachronic rival will rule out that PU has accept-and-write as an option, because PU is not certain she can diachronically accept-and-write. Any diachronic rival will say that PU has only minimal outer actions as options because these are the only outer actions that she is certain that she can diachronically do. Given this, any diachronic rival will say that PU ought to reject the invitation. Furthermore, it looks like any diachronic rival will entail that PP has accept-and-write as an option. If any diachronic rival rules *in* accept-and-write for PP, then any diachronic rival will entail that PP ought to accept and write, because *ex hypothesi* PP thinks that this leads to the best outcome. So any diachronic rival will entail that PU ought to reject, and that PP ought to accept-and-write. This creates a weird asymmetry about the relevance of the agent's predicted future irrationality. So any diachronic rival should be rejected.

There is one wrinkle here. More precisely, it looks like any *interesting* diachronic rival will entail that PP has accept-and-write as an option. For instance, a diachronic rival might say that an agent's options are *minimal* outer actions that she can diachronically do. This would rule out accept-and-write as an option for PP. However, this is *uninteresting* in the sense that this account is equivalent to one phrased in terms of a synchronic "can". It is equivalent to the account that says an agent's options are the minimal outer actions that she can synchronically do. Therefore, given that I am dealing with synchronic rivals later, there's no harm in ruling out at this stage all the uninteresting diachronic rivals along with the interesting diachronic rivals. So the conclusion stands: *any* diachronic rival should be rejected. In light of this, henceforth, any references to abilities will always be references to *synchronic abilities* unless otherwise stated.

3. Sobel's Objections to Rivals

In this section, I'm going to lay out Sobel's objections to three rivals to **Subjective Decisions** (Sobel 1983). Now, I don't agree with everything here – and I'll flag where I disagree. But I'm going to present Sobel's arguments first (subsections 3.1 to 3.4), and then evaluate their significance for my argument for **Subjective Decisions** (subsection 3.5).[40]

3.1 An Additional Constraint on Options: No Entailment

Before getting to Sobel's objections, I'm going to motivate an additional constraint on options. Many discussions of options assume that the option set satisfies some sort of structural constraint over the options.[41] Sobel also assumes a structural constraint (p.202). A structural constraint cuts down on the rivals to **Subjective Decisions**, and for that reason is attractive for a proponent of the latter. I *don't* think there is a structural constraint, which I'll come back to in subsection 3.5, but it is essential to Sobel's argument that there is one. So I will assume the following structural constraint for the purposes of outlining Sobel's arguments.

> No Entailment (henceforth *NE*). For all agents S, for any two of S's options, A and B, it is not the case that A *performance entails* B (relative to agent S).

("A" and "B", here and below, range over actions.) An option A *performance entails* another option B (relative to an agent S) just when S lacks the ability to perform A without performing B. For instance, an unskilled typist might be unable to type without looking at the keyboard, whereas a skilled typist might be able to do this.

[40] I should add that Sobel sees himself as arguing for **Subjective Decisions** by the same method as myself, i.e. ruling out rivals. I don't like Sobel's argument as it stands, because it fails to consider some notable rivals, for instance, diachronic rivals. Also, it doesn't consider a further rival discussed in section 4.

[41] See for instance Hedden (2012, p.347), Lewis (1981, p.7); Joyce (1999, p.52), Meacham (2010, p.51), Sobel (1983). The structural constraint is often put as a demand for mutual exclusivity between options. I think the difference between such a constraint and NE doesn't matter for the purposes of this chapter.

For the unskilled typist, *typing* performance entails *looking at the keyboard*, but it doesn't for the skilled typist. Henceforth I'll refer to *performance entailment* simply as *entailment*.

Why hold NE? Sobel gives no motivation for his structural constraint, but I think the standard argument goes as follows. Suppose (for *reductio*) that NE is false. Then there is an agent whose options include A and B such that A entails B. Now it could be that A is best, and, in particular, better than B. Given this, decision theory will entail that the agent ought to A, and that she ought not B. But this can't be right, because A entails B. Hence NE.[42]

For instance, suppose Jill walks into a clothes shop. Suppose she is certain that she can buy jumpers of various colours. Then you might think that her options are: buy a red jumper, buy a black jumper, buy a grey jumper, and so on. You might also think, in violation of NE, that one of her options is *buy a jumper*. Now suppose Jill really wants a *red* jumper – other coloured jumpers won't do. Then decision theory will say that Jill ought to buy a red jumper and that she ought not buy a jumper – but buying a red jumper entails buying a jumper! The idea behind NE is that something has gone wrong here, namely, allowing both *buy a red jumper* and *buy a jumper* to be options.

More generally, when the following are both options – first, performing an action, and, second, performing that action *in a specific way* – then NE is violated, because the latter action entails the former action. This is what happens in the case above. NE can be violated in at least two other ways as well.

First, suppose that *performing A at t1* is an option, and *performing A at t1 and B at t2* is also an option, where "t1" and "t2" label times. Then you have a violation of NE, because the latter action entails the former action. For instance, suppose that *raising your arm at t1* is an option, and also suppose that *raising your arm at t1 and again at t2* is an option. Then the latter entails the former, so NE is violated.

Second, when the following are both options: first, performing an action, and second, performing that action *plus more distinct actions at the same time*. For instance,

[42] This argument is discussed in the moral literature by Bergström (1968), Prawitz (1968) (1970), and Carlson (1997). Weirich (2001, pp.83-4) discusses it in relation to decision theory.

suppose raising your right arm is an option, and raising both arms is an option. Then the latter entails the former, so NE is violated.

One final thing about NE: note that the same motivation applies when the options are construed as *intentional* actions. The proponent of NE says that we don't want decision theory saying "agent ought A" and "agent ought not B" where A entails B – she says this regardless of whether A and B stand for actions or *intentional actions*. For instance, suppose you have the following two options: *intentionally raise your arm at t1* and *intentionally raise your arm at t1 and again at t2*. The proponent of NE says that we don't want decision theory entailing that you ought *intentionally raise your arm at t1 and t2* but that you ought not *intentionally raise your arm at t1*. For *intentionally raise your arm at t1 and t2* entails *intentionally raise your arm at t1*.

As I said above, NE is attractive to a proponent of **Subjective Decisions** because it immediately rules out some *prima facie* attractive rivals to **Subjective Decisions**. For instance, it immediately rules out the following:

> **All Actions.** Options are all and only the outer actions that the agent is certain she can do.

This account is ruled out by NE because there are cases where an agent is certain she can perform two actions such that one entails the other. For instance, as in Jill's case above, where she is sure that she can buy a jumper and sure that she can buy a red jumper.

Does **Subjective Decisions** satisfy NE? On a reasonably natural understanding of decisions, it appears it does. For suppose we understand a decision such that when an agent decides on A at a time, then A is *everything* that she decides on at that time. Then it's impossible to make two distinct decisions at the same time. For if an agent allegedly makes two distinct decisions at the same time, then they have different contents. But then neither decision is such that the agent has decided only on *that* decision's content.[43] In the next three subsections I'll look at Sobel's objections to three accounts that satisfy NE.

3.2 Against **NE Actions**

The first rival Sobel rejects is:

> **NE Actions**. An option set is *any* maximal set of outer actions satisfying NE such that the agent is certain she can perform each member.

A *maximal* set is a set such that no action could be added to the set to form a further option set. I'll go through why *I* reject **NE Actions** here. (I'm not convinced by Sobel's own argument, as I'll indicate below.)

I think **NE Actions** should be rejected because of the following case:

> Nim. The situation as it appears to Nim is as follows. Nim is driving and stuck in a congested lane. He will in fact change lanes. Given that he will change lanes, it is best that he accelerates, because otherwise he will slow down the speed of that lane. If he were to stay in his lane (contrary to what he will actually do), then it is best not to accelerate, otherwise he will plough into the truck in front of him. The best thing Nim could do is to stay put and not accelerate, because it is illegal to change lanes. This is the best thing that he could do despite the fact that he won't in fact do this: he will in fact change lanes.[44]

If **NE Actions** is right, then it looks like Nim's option sets include the following:

> {accelerate, don't accelerate}

This looks like an option set according to **NE Actions** because, first, it clearly satisfies NE. Second, Nim is certain he can do each member. And third, the set is maximal – addition of any other outer action Nim is certain he can do will mean it

[43] I remain neutral about this understanding of decisions. I think NE is not well-motivated (see below), so I don't have to endorse this conception of decisions.

[44] This case is a variant of one discussed in Goldman (1978).

fails to satisfy NE. For instance, if we add *accelerate-and-stay-put*, then this will entail *accelerate*. Given this option set, decision theory entails that Nim ought to accelerate. That's because he thinks he will in fact change lanes, and, given this, accelerating is best. However, this seems wrong – it's not the case that Nim ought to accelerate.

Now, if **NE Actions** is right, then it looks like Nim's option sets also include the following:

> {change-lanes-and-accelerate, change-lanes-and-don't-accelerate, stay-put-and-accelerate, stay-put-and-don't-accelerate}

This looks like an option set according to **NE Actions** because, first, it clearly satisfies NE. Second, Nim is certain he can do each member. And third, the set is maximal – addition of any other outer action Nim is certain he can do will mean it fails to satisfy NE. For instance, if we add *accelerate*, then this will be entailed by *stay-put-and-accelerate*. Given this option set, decision theory will say that Nim ought to stay-put-and-not-accelerate (because *ex hypothesi* Nim thinks that this leads to the best outcome). This seems like the right verdict. However, **NE Actions** delivers, along with this verdict, the verdict that Nim ought to accelerate, which seems wrong. The problem here is not that the two prescriptions conflict. The problem here is that **NE actions** entails something that's false, namely, that Nim ought to accelerate. For this reason I think **NE Actions** should be rejected.

In contrast, **Subjective Decisions** says that Nim's options are all and only the decisions that he is certain he can make. So his options include decisions for the more specific actions. Assuming Nim thinks his decisions would be efficacious, then *decide to stay-put-and-don't-accelerate* will receive the best EU, because this leads to staying put and keeping a constant speed, which *ex hypothesi* Nim thinks leads to the best outcome.[45]

[45] Both Hedden (2012) and Sobel (1983), discussing similar cases, object on the basis that **NE Actions** delivers conflicting prescriptions. I am not immediately concerned by conflicting prescriptions here because they are *relative to different option sets*. For instance, see Jackson & Pargetter (1986) and Jackson (2014) for an endorsement of conflicting prescriptions out of different option sets.

3.3 Against **Fully Specific Actions**

Now I move onto the second rival that Sobel considers. Nim's case suggests that if there is a plausible rival to **Subjective Decisions**, then it will entail that Nim's options are *more specific* ways of acting. That's because the option set containing the more specific actions, like *stay-put-and-don't-accelerate*, delivered the correct verdict rather than the option set containing the less specific actions, like *accelerate*. This suggests an account according to which the options are all and only the *fully specific* outer actions that the agent is certain she can do. More precisely:

> **Fully Specific Actions**. An action A is an option iff the following conditions are met:
>
> (i) *Certainty*. The agent is certain she can A.
>
> (ii) *Outerness*. A is an outer action.
>
> (iii) *Specificity*. There is no distinct action B such that: (a) B entails A, and (b) B satisfies conditions (i) and (ii).

Specificity ensures that any option is not entailed by a distinct option, so this account satisfies NE. **Fully Specific Actions** delivers the following options in Nim's case:

change-lanes-and-accelerate
change-lanes-and-don't-accelerate
stay-put-and-accelerate
stay-put-and-don't-accelerate

These are the fully specific outer actions that Nim can do. *Accelerate* and *don't accelerate* aren't options because they fail *Specificity*. For instance, *accelerate* is entailed by *stay-put-and-accelerate*, which Nim is certain he can do. Given this option set, decision theory says that Nim ought to stay-put-and-not-accelerate, because *ex hypothesi* Nim thinks that this action leads to the best outcome. This seems right.

However, Sobel objects to **Fully Specific Actions**. His objection assumes the CDT framework – that is, it assumes that the relevant evaluation of options is a *causal* evaluation. I'm happy to make this assumption – I think an account of options should be adequate in both EDT and CDT frameworks. So, in what follows, EU's should be understood as *causal* EU's. (Most of time, however, it won't matter whether the EU's are interpreted as evidential or causal EU's.) Sobel's objection to **Fully Specific Actions** is that it delivers a strange prescription in the following case: (N.B. Sobel treats "A is open for agent" and "agent can A" as synonymous.)

> [*Levers 1*.] Let there be two levers, *a* and *b*. Assume that the agent is sure that he can at no time pull both *a* and *b*, and at each of t1 and t2 must pull one of *a* and *b*. Let A1, his pulling *a* at t1, be "now" (i.e., just before t1) certainly open for him. And assume that he is sure that by t1 he can at will not only make certain A1, but also either A2 (his pulling *a* again at t2) or B2 (his pulling *b* at t2). Assume that not only the "immediate action" A1 but also the "extended actions" (A1&A2) and (A1&B2) are certainly open just before t1. Our agent is sure that he has the capacity (the self-control) not only to settle by t1 what he will do then, but also to commit himself in relation to t2 and settle by t1 what actions with first moment t2 he will do. However, let him be sure that, though he can settle what he does at t2, if he decides now for A1 he can leave *open* what he will do at t2. But matters are not the same if he decides now for B1; we assume that he is sure that pulling *b* would render *a* inoperable, so that in making B1 certain he would make B2 certain as well… We assume that the agent thinks that only two possible patterns of payoffs are at all likely, and that he thinks these patterns are equally likely. These patterns are: $0 each time *a* is pulled, and $3 each time *b* is pulled; and $2 each time *a* is pulled, and $0 each time *b* is pulled. (Sobel 1983, p209)

So the basic set-up is that there are two levers, exactly one of which must be pulled at two different times, t1 and t2 (where t1 is earlier than t2). A1 labels the action of pulling *a* at t1, etc. The agent is certain that she can – in the synchronic sense – A1, A1&A2, A1&B2, B1, and B1&B2. That is, as Sobel puts it, the agent is sure she can "make certain" or "commit to" all of these actions, both minimal and extended (outer) actions. I will also adopt this terminology. I will assume (something that is implicit in the quoted passage) that the agent is sure she can commit herself to the various minimal and extended (outer) actions by making *decisions* to pull the various levers. For instance, I assume that the agent is sure that she can commit herself to A1&A2 because she is sure that she can decide on A1&A2, and she is sure that this decision would ensure that she performs A1&A2. (In the remainder of this section on Sobel, whenever I talk of the minimal and extended actions I mean the minimal and extended *outer* actions, but for convenience I drop "outer".)

There is a final detail to this case. Sobel assumes the following: the agent is sure that if she were to make certain only A1, then she would pull whichever lever is most lucrative at t2 – she will know at t2 which lever is most lucrative because she'll be able to tell which lever has the prize from the results of her first pulling (p.210). Given this, the intuition is that the agent ought to make certain only A1, leaving her options open at t2 so she can exploit the incoming information (something she is certain she would do). The question now is: does **Fully Specific Actions** deliver a verdict that respects this intuition?

Before getting on to what **Fully Specific Actions** says, it's worth looking at how **Subjective Decisions** treats this case. **Subjective Decisions** says that the options are all and only the decisions that the agent is certain she can make. Let "d(A)" stand for a decision for A.[46] Then the contenders for being best are below (followed in each case by what I assume the agent is sure the decision makes certain):

d(A1) – makes certain only A1

d(B1) – makes certain B1&B2

d(A1&A2) – makes certain A1&A2

d(A1&B2) – makes certain A1&B2

d(B1&A2) – makes certain B1&B2

d(B1&B2) – makes certain B1&B2

As Sobel explicitly stipulates, the agent is sure that if she were to make certain only A1, then she would pull whichever lever is most lucrative at t2 (p.210). Given this, the EU's of these decisions are as follows (where financial gain lines up with utility):

EU(d(A1)) = 0.5*2 + (0.5*3 + 0.5*2) = 3.5

EU(d(B1)) = EU(d(B1&A2)) = EU(d(B1&B2)) = 0.5*3 + 0.5*3 = 3

EU(d(A1&A2)) = 0.5*2 + 0.5*2 = 2

[46] I will continue to use "A" as ranging over actions, even though there is the risk of confusion with A1.

EU(d(A1&B2)) = 0.5*2 + 0.5*3 = 2.5

For instance, the agent is sure that d(A1) would make certain only A1. This has half a chance of delivering $2 at t1 and then, because the agent is sure that she would pull the most lucrative lever at t2, half a chance of delivering $2 at t2, and half a chance of delivering $3 at t2. So it has an EU of 3.5. Given these EU's, the decision for A1 will come out best. So on **Subjective Decisions**, the agent ought to decide A1. This respects the intuition that the agent ought to make certain only A1.

In contrast, **Fully Specific Actions** says that the agent's options are the fully specific outer actions that she is certain she can do. So it says her options are:

A1&A2

A1&B2

B1&B2

B1&A2 is not an option because the agent is sure that pulling *b* at *t1* makes *a* inoperable, so the agent is sure she can't do that. Neither A1 nor B1 are options because they are not *fully specific* outer actions that the agent is certain she can do.

Given that the options are these extended actions, decision theory will entail that the agent ought to B1&B2. Why? Well consider how we evaluate these extended actions *in the CDT framework*. The agent keeps her views about the causal structure of the world fixed and evaluates what the upshot would be if she were to perform these actions. So because the agent thinks the prize is as likely to result from *a* as *b*, then the agent thinks A1&A2 is 0.5 likely to result in no money at all, and 0.5 likely to result in $4. So assuming financial gain lines up with utility, it has a causal EU of 2. By the same reasoning, the causal EU's of these actions are as follows:

EU(A1&A2) = 0.5*0 + 0.5*4 = 2

EU(A1&B2) = 0.5*3 + 0.5*2 = 2.5

EU(B1&B2) = 0.5*6 + 0.5*0 = 3

So decision theory says that the agent ought to B1&B2.

Intuitively, it's not the case that the agent ought to B1&B2. The intuition is that the agent ought to make certain only A1. What's gone wrong for **Fully Specific Actions** here? I think that there are many ways of saying what's gone wrong. Sobel says the problem is that "in some case where it would be rational for an agent to leave future options open, this principle selects actions as rational that would close them" (p.210). The idea is that **Fully Specific Actions** entails that the agent ought to pull lever *b* at t1, which means she makes *a* inoperable at t2, thus wrongly closing off her options at t2. In contrast, I think it's useful to compare **Fully Specific Actions'** treatment of this case to **Subjective Decision's** treatment. Note that in this case, the extended actions receive the same causal EU as *decisions for* the extended actions. So the extended actions can be thought of as surrogates for the decisions for the extended actions. Given this way of looking at it, **Fully Specific Actions** goes wrong because it completely ignores the agent's decisions for minimal actions. This is a problem because, intuitively, the agent ought to make a decision for a certain minimal action. So it's no wonder that **Fully Specific Actions** gets this case wrong.[47]

3.4 Against **Minimal Fully Specific Actions**

Sobel says that the failure of **Fully Specific Actions** in *Levers 1* suggests that the decisive evaluations should be of the minimal actions, A1 and B1, rather than the extended actions, A1&A2, A1&B2, etc. This suggests an account, according to which, the options are all and only the *minimal* fully specific outer actions that the agent is certain she can do. More precisely:

> **Minimal Fully Specific Actions**. An action A is an option iff the following conditions are met:

[47] There may be another objection to **Fully Specific Actions**, namely, that because it entails that outer actions are options, decision theory will deliver the verdict that the agent ought to perform a certain outer action. But the intuition is at the level of what the agent ought to *make certain*. It seems that **Fully Specific Actions** makes decision theory's prescription a prescription to do the wrong *sort* of thing. I think this is right. However, I note that a proponent of **Fully Specific Actions** might reject my deflationary reading of "choose" so that the prescription is that the agent ought to *choose* a specific outer action. Given this, it's not obvious that the prescription is pitched at the wrong level, for it will depend on one's account of choice.

(i) *Certainty*. The agent is certain she can A.

(ii) *Outerness*. A is an outer action.

(iii) *Minimality*. A is a minimal action.

(iv) *Fully Specific*. There is no distinct action B such that: (a) B entails A, and (b) B satisfies conditions (i), (ii), and (iii).

In *Levers 1*, this account entails that the options are:

A1

B1

These are the minimal fully specific outer actions that the agent is certain she can do. None of the extended actions are options because they are not *minimal*. Sobel says that A1 would receive the best evaluation. He assumes that the agent is sure that if she were to A1, then she'd leave her options open at t2 and pull whichever lever is most lucrative (p.211). Given this assumption, the agent thinks that if she were to A1, then half of the time she'd win $4 and the other half she'd win $3, so it has an EU of 3.5. In contrast, the agent is sure that if she were to B1, then she'd B2 (because pulling *b* makes *a* inoperable). So the agent thinks that if she were to B1, then half the time she'd win $6, and the other half she'd win nothing, hence B1 has an EU of 3. Thus decision theory would say that the agent ought to A1. This seems like a sensible verdict. (As I said, **Minimal Fully Specific Actions** gets *Levers 1* right *given* the assumption that the agent is sure that if she were to A1, then she'd leave her options open at t2 and pick whatever lever is most lucrative – we'll come back to this assumption later.)

What's happened here is that, as Sobel sets things up, the minimal actions have the same EU as *decisions for* the minimal actions. So the minimal actions can be thought of as surrogates for the decisions for the minimal actions. I said above that **Fully Specific Actions** goes wrong because it ignores the agent's decisions for minimal actions. **Minimal Fully Specific Actions** doesn't have that problem, and so it delivers a verdict that respects the intuition.

However, Sobel has an objection to **Minimal Fully Specific Actions**. Essentially, the problem is that it ignores the agent's decisions for extended actions. Although this doesn't cause trouble in *Levers 1*, it does in the following case:

> [*Levers 2*.] The agent is sure that the set-up is as follows. There are two levers. They cannot at any time both be depressed. Each can be depressed at t1, but each can be depressed at t2 if and only if one or the other was depressed at t1. Depressing *a* at t1 delivers $5. Depressing *b* at t1 delivers $1. After t1 and before t2, $10 is added either to the $5 that *a* delivers in any event at t2 or to the $1 that *b* delivers in any event at t2. The position of the $10 bonus is determined by a random device, so that the chance of its being associated with *a* is the same as that of its being associated with *b* (this chance is thus .5). If *b* is depressed at t1, then shortly thereafter the position of the $10 bonus is announced by a clear signal that the agent would interpret correctly and be able successfully to act upon at t2. But if *a* is depressed at t1 then – though the position of the bonus is announced by a signal – the signal is made so late, is so unclear, and is so brief that it is only as likely as not that the agent correctly interprets it and successfully acts upon it.
>
> So much for the set-up. We now proceed to several of the agent's beliefs concerning actions. A1 and B1 are to be fully specific among minimal actions that are certainly open just before t1. Extended actions (A1&A2) and (A1&B2) are certainly open just before t1, but he is sure they are not minimal; and (B1&A2) and (B1&B2) are not certainly open just before t1. Furthermore, the agent is sure that if he were to opt for A1 then, though he *could* incorporate its choice in the choice of a more extended action, he would not. He is sure that if he were to opt for A1 he would gamble (unwisely) on the unclear signal and try to act on it. In contrast, he is sure that if he were to opt for B1 then he not only would not but could not incorporate its choice in the choice of a more extended action: He is sure that, were he to do B1, then whatever his state of mind at t1 – whatever his "resolve" – he would at t2 be completely under the sway of the clear signal that would have announced the position of the $10 bonus. (Sobel 1983, pp.211-212)

Sobel says that the intuition here is that the agent ought to commit herself to pulling *a* twice. This is not obvious. So let's go through how **Subjective Decisions** treats this case, after which the intuition will become a bit clearer. **Subjective Decisions** says that the options are all and only the decisions that the agent is certain she can make. The contenders for being best are below (followed in each case by what I assume the agent is sure they would make certain):

d(A1) – makes certain only A1

d(B1) – makes certain only B1

d(A1&A2) – makes certain A1&A2

d(A1&B2) – makes certain A1&B2

d(B1&A2) – makes certain only B1

d(B1&B2) – makes certain only B1

As Sobel explicitly stipulates, the agent is sure that if she makes certain only A1, then she will unwisely guess the location of the prize at t2. And she is sure that if she makes certain only B1, then she will be under the sway of the signal. Given this, the EU's of these decisions are:

EU(d(A1)) = 5 + (0.25*15 + (0.25*5 + (0.25*11 + 0.25*1))) = 13

EU(d(B1)) = EU(d(B1&B2)) = EU(d(B1&A2)) = 1 + (0.5*11 + 0.5*15) = 14

EU(d(A1&A2)) = 5 + (0.5*15 + 0.5*5) = 15

EU(d(A1&B2)) = 5 + (0.5*11 + 0.5*1) = 11

For instance, if the agent performs d(A1), then she gets $5 from the first pulling, and then she will guess the location of the prize. Thus she has a quarter of a chance of all of the following – correctly guessing *a*, wrongly guessing *a*, correctly guessing *b*, and wrongly guessing *b* – and she would win $15, $5, $11, and $1 respectively. Hence d(A1) has an EU of 13. Given these EU's, the decision for A1&A2 will come out best. This seems like a sensible verdict. It seems that the agent ought to decide on A1&A2, or make certain A1&A2, because that would guarantee her $10 from pulling *a* twice and she would have half a chance of getting the $10. Moreover, it would stop her unwisely guessing the location of the prize at t2 (as she would do if she made certain only A1).

Now let's look at what is entailed by **Minimal Fully Specific Actions**. As Sobel says in the quoted passage, on **Minimal Fully Specific Actions**, the options are A1 and B1, because these are the minimal fully specific outer actions that the agent is sure she can do. Given these options, decision theory will recommend B1. Why? As Sobel stipulates, the agent is sure that if she were to first pull *a*, then she would leave her options open at t2. And she is sure that if she were to first pull *b*, then she would be under the sway of the signal. Given this, the EU of A1 is the EU of pulling *a* first and then unwisely guessing the location of the $10. So it is: 5+(0.25*15+(0.25*11+(0.25*5+0.25*1)))=13. In contrast, the EU of B1 is the EU of pulling *b* first and then being under the sway of the signal. So it is: 1+(0.5*11+0.5*15)=14. So decision theory recommends B1.

Intuitively, it's not the case that the agent ought to B1. Intuitively, the agent ought to make certain A1&A2. What's gone wrong for **Minimal Fully Specific Actions** here? I think that there are many ways of saying what's gone wrong. Sobel says that "It can select actions that would leave open options that ought to be closed" (p.211). The idea is that **Minimal Fully Specific Actions** entails that the agent ought to B1. If she were to B1, she would not settle how things go at t2. But intuitively, she should settle how things go at t2 by making certain A1&A2. In contrast, I think it's useful to contrast **Minimal Fully Specific Actions'** treatment with **Subjective Decision's** treatment. In this case, the minimal actions receive the same EU as *decisions for* the minimal actions. That's because Sobel assumes that if the agent were to perform A1, then she would gamble unwisely on the unclear signal; and if the agent were to perform B1, then she would be under the sway of the clear signal at t2. This is exactly what would happen if the agent were to decide on A1 and decide on B1 respectively. So the minimal actions receive the same EU as decisions for the minimal actions. Given this, the minimal actions can be thought of as surrogates for the *decisions for* the minimal actions. On this way of looking at it, **Minimal Fully Specific Actions** goes wrong because it completely ignores the agent's decisions for extended actions. Ignoring such decisions is problematic because, in this case, it is intuitive that the agent ought to make a decision for a certain extended action. So it's no wonder that **Minimal Fully Specific Actions** goes wrong in this case.

That completes Sobel's objections to **Fully Specific Actions** and **Minimal Fully Specific Actions**. In summary, the two cases are constructed such that an outer action receives the same causal EU as a decision for the action. **Fully Specific Actions** entails that only the extended actions in each case are the options. This means it effectively ignores the agent's decisions for minimal actions, and so gets in trouble when the intuition is that the agent ought to make a decision for a certain minimal action, as in *Levers 1*. **Minimal Fully Specific Actions** faces the reverse problem. It says that only the minimal actions in each case are the options. This means it effectively ignores the agent's decisions for extended actions, and so gets in trouble when the intuition is that the agent ought to make a decision for a certain extended action, as in *Levers 2*.

Now you might think that one way of solving these problems is to combine **Fully Specific Actions** and **Minimal Fully Specific Actions** and include both minimal and extended actions in the option set. The problem with this is that it *doesn't* satisfy NE. That's because the extended actions entail the minimal actions. In other words, if we want to get the right verdicts in these two cases and have outer actions as options, then the option set needs to include *all* outer actions, both minimal and extended. However, NE effectively forces us to choose between minimal and outer actions, so there is no way of delivering the right verdicts with outer actions as options.

3.5 The Significance of Sobel's Arguments

In objecting to **NE Actions**, **Fully Specific Actions**, and **Minimal Fully Specific Actions**, Sobel's considers himself to have objected to all (plausible) rivals to **Subjective Decisions**. So he concludes that **Subjective Decisions** is the genuine account of options. You might think that along with my rejection of diachronic rivals (see section 2), this is a good argument for **Subjective Decisions**.

However, I think NE is not well-motivated, as I'll now explain. Recall the motivation for NE. Suppose (for *reductio*) that NE fails. Then there is an agent whose options include A and B such that A entails B. Now it could be that A is best, and, in particular, better than B. Given this, decision theory will say that the agent ought to A, and that she ought not B. But this can't be right, because A entails B. Hence NE.

Note that this argument assumes that decision theory entails that an agent ought to realise an option that is best and that she ought not realise an option that isn't best. One way to avoid decision theory delivering these conflicting prescriptions is to change how decision theory generates verdicts about what the agent ought do and ought not do. For instance, one could formulate decision theory so that it comprises of the following three claims:

- An agent ought to realise an option if it is the best option *or if it is entailed by the best option*;
- It is permissible for an agent to realise an option if it is maximal *or if it is entailed by a maximal option*;[48]
- An agent ought not realise an option if it is neither obligatory nor permissible.[49]

Given this, decision theory would avoid generating conflicting prescriptions. So there is a way to respect the motivation for NE without upholding it as a genuine constraint. My point here is that NE is *not* well-motivated in the present context. More specifically, given that there is another way to respond to the motivation for NE, and I don't have any argument to prefer NE over reformulating decision theory, it is incumbent on me to consider rivals to **Subjective Decisions** which *do not* satisfy NE.[50]

Where does all of this leave Sobel's objections and my argument for **Subjective Decisions**? Sobel's objections to the three rivals to **Subjective Decisions** remain objections to those three rivals. I think we should see Sobel's argument as demonstrating what goes wrong when you try to reduce the option set to particular sorts of outer actions – extended actions or minimal actions, for instance. But to

[48] Recall from Chapter 1 (section 3) that an option is best when its EU is *the* greatest; an option is maximal when its EU is no worse than any other option's EU.

[49] I note that Weirich (2001, pp.83-4) alludes to a similar reformulation of decision theory in his discussion of a structural constraint on the option set.

[50] In the previous chapter I endorsed a change to the standard formulation of decision theory. I suggested decision theory says that an agent ought to do *as much as she can* of the best option. I've just considered a *further* change to the standard formulation of decision theory. It might be that it is difficult to combine these two suggested changes into one plausible formulation. However, this would only strengthen my hand, because then there would be reason to uphold NE, and then there would be fewer rivals to **Subjective Decisions** to consider.

obtain a comprehensive argument for **Subjective Decisions**, one must consider rivals that do not satisfy NE. So a derivative conclusion of this chapter is that Sobel's argument does not establish **Subjective Decisions** because he's failed to consider all of the rivals to that account. I will consider a rival that does not satisfy NE in the next section. This rival is motivated by the failure of **Fully Specific Actions** and **Minimal Fully Specific Actions.** More specifically, their failures suggest that a satisfactory rival to **Subjective Decisions** would avoid reducing the option set to some species of outer action. So in the next section I will consider the following account of options:

> **All Actions**. Options are all and only the outer actions that the agent is certain she can do.

4. Against All Actions

The problem with **All Actions** is that outer actions are *too coarse-grained* to serve as options. Outer actions are compatible with different decisions. For instance, an agent's performing A1 is compatible with her deciding A1 and it is also compatible with her deciding A1&A2. This affects the evaluation of A1 and leads to trouble in certain cases. I will show this with two cases. The first assumes an EDT framework and focuses on extended actions. The second assumes a CDT framework and focuses on minimal actions. (As in the last section, when I talk of minimal and extended actions in this section, strictly-speaking, I mean minimal and extended *outer* actions.)

First, consider the following case:

> *Levers 3*. The situation as it presents itself to Harry is as follows. Suppose Harry has two levers in front of him, *a* and *b*, and he must pull exactly one at each of two times, t1 and t2. Pulling *b* at t1 makes *a* inoperable at t2. Pulling *b* is the safe option – if Harry pulls it twice he neither wins nor loses

> anything. Pulling *a* at t1 is risky. If he pulls *a* at t1 then a random generator will determine which of the following scenarios is in play for t2:
>
> > S1. Good Scenario. Pulling *a* at t2 delivers $100; pulling *b* at t2 delivers $0.
> >
> > S2. Bad Scenario. Pulling *a* at t2 means losing $240; pulling *b* at t2 means losing $230.
>
> So pulling *a* is best if the Good Scenario obtains, but pulling *b* is best if the Bad Scenario obtains. Harry is told after t1 which is the situation, and each situation has a 0.5 probability. Harry can A1&A2, A1&B2, B1&B2, A1, and B1. That is, he can commit himself (by making the relevant decisions) to any combination of lever pullings (except for B1&A2, of course, because pulling *b* at t1 makes *a* inoperable at t2). That is the situation as it presents itself to Harry. Additionally, Harry is sure that if he makes certain only A1, then he pulls whichever lever is best to pull at t2 (i.e. *a* in the Good Scenario and *b* in the Bad Scenario). He will be able to do this because he is told which of the scenarios obtain after pulling *a* at t1.

Intuitively Harry ought to go for the safe option and pull *b* twice. Pulling *a* at t1 in the hope of getting the Good Scenario is not worth the risk. In the Good Scenario the most you can win is $100, but you will lose over $200 in the Bad Scenario.

Before looking at what **All Actions** entails in this case, let's see what is entailed by **Subjective Decisions**. This account says that Harry's options are all and only the decisions that he is certain he can make. The contenders for being best are below (followed in each case by what I assume the agent is sure they would make certain):

d(A1) – makes certain only A1
d(B1) – makes certain B1&B2
d(A1&A2) – makes certain A1&A2
d(A1&B2) – makes certain A1&B2
d(B1&A2) – makes certain B1&B2
d(B1&B2) – makes certain B1&B2

Now, Harry is sure that in making certain only A1, he would choose whichever lever is most lucrative at t2 (depending on whether Good Scenario or Bad Scenario obtains). Given this, the EU's are as follows:

EU(d(A1)) = 0.5*100 + 0.5*(-230) = -65

EU(d(B1)) = EU(d(B1&B2)) = EU(d(B1&A2)) = 0

EU(d(A1&A2)) = 0.5*100 + 0.5*(-240) = -70

EU(d(A1&B2)) = 0.5*0 + 0.5*(-230) = -115

So decision theory would entail that it is permissible for Harry to decide on B1, B1&B2, or B1&A2 – a sensible verdict, given that all of these decisions would mean that he pulls *b* twice.

Now consider what is entailed by **All Actions**. In particular A1&A2 is in the option set – because that is something that Harry is certain he can do. Now one way of performing A1&A2 is by deciding on A1&A2 and thus making A1&A2 certain. The problem is that there is another way to perform A1&A2, namely, by first deciding to A1 (and thus making certain only A1), and then subsequently performing A2 (without having committed to it earlier). This should already give us pause for thought – because this doesn't seem like a natural way of carving up an option. Regardless, it leads to trouble.

To see this, suppose that Harry is certain that if he performs A1, then he makes certain only A1 (that is, suppose his conditional credence for making certain only A1 given that he performs A1 is 1).[51] Now, assuming an evidential evaluation, the extended action A1&A2 receives the best evaluation. That's because: any (doxastically) possible world in which Harry performs A1&A2 is, of course, a world in which he performs A1. I've just stipulated that Harry is sure that if he performs A1, then he makes certain only A1. So any world in which Harry performs A1&A2,

[51] This supposition doesn't affect the EU's of the decisions calculated earlier *as long as those EU's are well-defined.* The supposition entails that Harry is sure that he *won't* make certain A1&A2 nor A1&B2, and so sure that he won't make the corresponding decisions. On some EDT's this would mean the decisions A1&A2 and A1&B2 don't have well-defined EU's. I assume EU's are still well-defined. Regardless, I make a different supposition to demonstrate the same problem below, and this *doesn't* endanger the EU's of the decisions calculated earlier.

is a world in which he makes certain only A1. I set up the case earlier such that Harry is sure that if he makes certain only A1, then he pulls whichever lever is most lucrative at t2. So any world in which Harry performs A1&A2, is a world in which he makes certain only A1 and then it is most lucrative to pull *a* at t2. This happens when the Good Scenario obtains. So any world in which Harry performs A1&A2, is a world in which the Good Scenario obtains. So any world in which Harry performs A1&A2, he wins $100. Thus the EU of A1&A2 is 100. This is the best outcome *ex hypothesi,* hence decision theory will say that Harry ought to A1&A2. This is the wrong verdict.

Now you might think that there is something strange about Harry being sure that if he performs A1, then he makes only A1 certain. In fact, this isn't so strange, because if he is going to perform A1, then the most rational thing to do is to make certain only A1 (so that he can at least pull whichever lever is best at t2, something he is certain he will do if he makes certain only A1). Regardless, I can allow that Harry is, say, 0.5 confident that he makes certain only A1 given that he performs A1, 0.25 confident that he makes certain A1&A2 given that he performs A1, and 0.25 confident that he makes certain A1&B2 given that he performs A1. Decision theory would still deliver the wrong verdict.

Why would decision theory still deliver the wrong verdict with the credences as above? The relevant credences are:

Cr(makes certain only A1|A1) = 0.5
Cr(makes certain A1&A2|A1) = 0.25
Cr(makes certain A1&B2|A1) = 0.25

Now let's try to work out the evidential EU of A1&A2. In half of the worlds where Harry makes certain only A1, the Good Scenario obtains, he performs A2, and he gets $100. In the other half, the Bad Scenario obtains, so he performs B2. So 25% of the A1-worlds are where he performs A2 and wins $100.

Similarly, in half of the worlds where he makes certain A1&A2, the Good Scenario obtains, so he gets $100. In the other half, the Bad Scenario obtains and he loses $240. So another 12.5% of the A1-worlds are worlds where Harry performs

A2 and wins $100. And 12.5% of the A1-worlds are where Harry performs A2 and loses $240.

In the worlds where he makes certain A1&B2 he doesn't A1&A2 so we don't need to consider such worlds. So 37.5% of the A1-worlds are where Harry performs A2 and wins $100. And 12.5% of the A1-worlds are where Harry performs A1 and loses $240. So in 75% of the A1-worlds where he also performs A2, Harry wins $100; and in 25% of the A1-worlds where he also performs A2, he loses $240. That means the evidential EU of A1&A2 is: (0.75*100)+(0.25*(-240))=15. This causes trouble. For the intuition is that the agent ought to pull *b* twice. So out of the options, only decision theory recommending B1 or B1&B2 would respect the intuition. But their EU's are 0, because pulling *b* first makes *a* inoperable, so they both result in pulling *b* twice, which is stipulated to win Harry nothing. Thus decision theory will not deliver the correct verdict here.

What seems to be going on is that A1&A2 is too coarse-grained – one way of performing it is by making certain A1&A2. The natural way of evaluating it is by supposing that the agent makes certain A1&A2. However, another way of performing it is by making certain only A1 and then performing A2 later. So as long as Harry assigns credence to performing it in this latter way, then its evaluation is sensitive to this way of performing it. This makes A1&A2 look really good, with the result that decision theory recommends it. But, intuitively, A1&A2 is irrational.[52]

Another case where the same problem surfaces for **All Actions** is in a tweak of Sobel's *Levers 1* case, and in a CDT framework. Here is the original case:

> [*Levers 1*.] Let there be two levers, *a* and *b*. Assume that the agent is sure that he can at no time pull both *a* and *b*, and at each of t1 and t2 must pull one of *a* and *b*. Let A1, his pulling *a* at t1, be "now" (i.e., just before t1) certainly open for him. And assume that he is sure that by t1 he can at will not only make certain A1, but also either A2 (his pulling *a* again at t2) or B2 (his pulling *b* at t2). Assume that not only the "immediate

[52] This objection also cuts against an account of options that says they are all and only the actions (*both* outer actions *and* decisions) that the agent is certain she can do. For this objection is a problem for any account that entails that A1&A2 is an option, which the account in question does.

> action" A1 but also the "extended actions" (A1&A2) and (A1&B2) are certainly open just before t1. Our agent is sure that he has the capacity (the self-control) not only to settle by t1 what he will do then, but also to commit himself in relation to t2 and settle by t1 what actions with first moment t2 he will do. However, let him be sure that, though he can settle what he does at t2, if he decides now for A1 he can leave *open* what he will do at t2. But matters are not the same if he decides now for B1; we assume that he is sure that pulling *b* would render *a* inoperable, so that in making B1 certain he would make B2 certain as well… We assume that the agent thinks that only two possible patterns of payoffs are at all likely, and that he thinks these patterns are equally likely. These patterns are: $0 each time *a* is pulled, and $3 each time *b* is pulled; and $2 each time *a* is pulled, and $0 each time *b* is pulled. (Sobel 1983, p.209)

There is a final detail to this case. Sobel assumes the following: the agent is sure that if she were to make certain only A1, then she would pull whichever lever is most lucrative at t2 – she will know at t2 which lever is most lucrative because she'll be able to tell where the prize is from the results of her first pulling (p.210). Given this, the intuition is that the agent ought to make certain only A1, leaving her options open at t2 so she can exploit the incoming information (something she is certain she would do).

We've already seen that **Subjective Decisions** delivers the correct verdict in this case, but let me quickly go through what it entails. I will assume (something that is implicit in the quoted passage) that the agent is sure she can commit herself to the various minimal and extended actions by making *decisions* to pull the various levers. For instance, I assume that the agent is sure that she can commit herself to only A1 because she is sure that she can decide on A1 and she is sure that this decision would ensure that she performs A1. Given this, the EU's of the decisions (that are contenders for having maximal EU) are:

$$EU(d(A1)) = 0.5*2 + (0.5*3 + 0.5*2) = 3.5$$

$$EU(d(B1)) = EU(d(B1\&B2)) = EU(d(B1\&A2)) = 0.5*3 + 0.5*3 = 3$$

$$EU(d(A1\&A2)) = 0.5*2 + 0.5*2 = 2$$

$$EU(d(A1\&B2)) = 0.5*2 + 0.5*3 = 2.5$$

For instance, the agent is sure that d(A1) would make certain only A1. This has half a chance of delivering $2 at t1 and then, because the agent is sure that she would pull the most lucrative lever at t2, half a chance of delivering $2 at t2, and half a chance of delivering $3 at t2. So it has an EU of 3.5. Hence the decision for A1 will come out best. Thus on **Subjective Decisions**, the agent ought to decide A1. This respects the intuition that the agent ought to make certain only A1.

The question is: what does **All Actions** entail? If options are outer actions, then we're aiming for a prescription that says the agent ought to A1. **All Actions** says that the agent's options are all and only the outer actions that she is certain she can do. The agent is certain she can A1, B1, A1&B2, A1&A2, B1&B2, so they are her options. At first, this looks promising, for the option set includes A1. The problem is that A1 is consistent with making certain only A1 and making certain both A1&A2. Now in Sobel's original case, he makes the following crucial assumption:

> (A) The agent is sure that if she were to A1, then she would keep her options open at t2 and would choose whatever is most lucrative at the later time.

This makes A1's evaluation the same as making certain only A1, with the result that A1 looks really good. The upshot is that decision theory recommends A1 – the right verdict.

However, note that (A) means that the agent is certain she *won't* commit herself to A1&A2. For in a world in which the agent commits herself to A1&A2, if she were to A1, then she *wouldn't* keep her options open at t2. Of course, it's Sobel's case, so there is nothing wrong with him setting it up however he wants. But it's natural to wonder what happens when this assumptions fails, so that the agent assigns some credence to committing herself to A1&A2. In such a case, **All Actions** is in trouble.

To see the problem, suppose the agent assigns 0.35 credence to committing herself to A1&A2. Then the best case for a proponent of **All Actions** is that in the other regions of doxastic space (i.e. where she doesn't commit herself to A1&A2), if the agent were to A1, then she would keep her options open at t2 and choose the most lucrative lever. In such a situation, the causal EU of A1 is an average of the EU's of committing herself to A1&A2 (which is 2) and making-certain-only-A1-and-then-choosing-whichever-lever-is-most-lucrative (which is 3.5), weighted by

0.35 and 0.65 respectively. In this case, the causal EU of A1 is slightly less than 3. The problem is that the causal EU of both B1 and B1&B2 is 3 (because if the agent were to pull *b* at t1, then she would pull *b* at t2, and she would have half a chance of obtaining $6). So decision theory would say that it is permissible for the agent to B1 and B1&B2 – the wrong verdict.

What's gone wrong here is that A1 is too coarse-grained. One way of performing it is by making certain only A1. That's the natural way to evaluate it. However, on reflection, there is another way of performing it, by making certain A1&A2. That means its EU is sensitive to worlds in which she makes certain this extended action. Making certain A1&A2 is bad in this case. The result is that A1 is evaluated badly, so decision theory won't recommend it. Thus decision theory doesn't capture the intuition on **All Actions**.

In summary, **All Actions** fails to deliver sensible verdicts in *Levers 3* and a tweaked *Levers 2*. The problem is that outer actions are too coarse-grained to serve as options. On the one hand, *Levers 3* exploits the fact that an outer action is compatible with making a decision now, and compatible with making one later contingent on some very good circumstance obtaining. Given this, the outer action looks good, because the goodness of the circumstance is being factored into the evaluation of the outer action. This results in the outer action being mistakenly recommended. On the other hand, the tweaked *Levers 2* case exploits the fact that performing an outer action A is compatible with different decisions and it seems we want to say that the agent ought to make a particular one of these decisions, call it *d*. If options are outer actions, then the only way to capture this intuition is by A being recommended. But if the non-d ways of performing A are sufficiently bad, then A won't be recommended. In both cases, it's the *coarseness* of outer actions which causes trouble: they are compatible with making different decisions that lead to them. So it is for this reason – the coarse-grainedness of outer actions – that I reject **All Actions**.

5. Conclusion

In this chapter I've rejected a number of rivals to **Subjective Decisions**, that is, a number of accounts that conceive of options as outer actions. First, I rejected any diachronic rival, that is, any rival that appeals to the diachronic – rather than the synchronic – "can". The problem here was that it led to strange asymmetries in the relevance of the agent's predicted future irrationality. That's because, on a diachronic "can", an agent can be certain that she can perform an extended outer action, and so it would count as an option on a diachronic rival, despite the fact that she predicts that she will irrationally abandon this action. This makes the agent's predicted irrationality irrelevant. In itself I have no objection to this. However, what's problematic is that the agent's predicted irrationality *becomes relevant* as soon as she starts to doubt her diachronic ability to perform the extended action. That's because, when she starts to have such doubts, her options shrink to only the minimal actions, whose decision theoretic evaluations take into account predicted irrationality. The upshot is a strange asymmetry in the relevance of the agent's predicted future irrationality. For that reason I reject diachronic rivals.

This set the stage for Sobel's discussion of the following three rivals (all employing the synchronic "can"):

- **NE Actions**. An option set is any maximal set of outer actions that satisfies NE each member of which the agent is certain she can do.
- **Fully Specific Actions**. Options are all and only the fully specific outer actions that the agent is certain she can do.
- **Minimal Fully Specific Actions**. Options are all and only the minimal fully specific outer actions that the agent is certain she can do.

I offered my own argument against **NE Actions**. The problem for **NE Actions** is that it leads to an intuitively false verdict in Nim's case. The problem is that it is too liberal in what counts as an option set, in particular, it allows that an option set of coarse-grained outer actions counts as an option set. This means that the agent will be recommended to perform some coarse-grained outer action because she will *in* *fact* perform the other coarse-grained outer actions in a very poor way (even though she has the ability to perform these other actions in a very good way). It seems that the decisive evaluations should be of the finer-grained outer actions. For that reason, I reject **NE Actions**.

Having rejected **NE Actions**, I outlined Sobel's objections to **Fully Specific Actions** and **Minimal Fully Specific Actions**. The problem for these two rivals is that if options are outer actions, then the option set should include *all* outer actions. However, **Fully Specific Actions** and **Minimal Fully Specific Actions** select a particular subset of all the outer actions, thus delivering intuitively false verdicts in some concrete cases. Finally, I objected to **All Actions**: options are all and only the outer actions that the agent is certain she can do. The problem here is that outer actions are too coarse-grained – they are compatible with making different decisions – thus resulting in some strange verdicts.

All five rivals – diachronic rivals, **NE Actions**, **Fully Specific Actions**, **Minimal Fully Specific Actions**, and **All Actions** – deliver intuitively false verdicts in concrete cases. In contrast, as I showed as I went along, **Subjective Decisions** gets all of these cases right. I conclude that **Subjective Decisions** is the genuine accounts of options. An agent's options are all and only the decisions that she is certain she can make. (As mentioned in section 1, this is made under the assumption that, for every decision, the agent is either certain she can make it or certain she can't make it. I'll be discussing cases where this assumption fails in the next chapter.)

This conception of options has a number of important consequences. First of all, it's in tension with our ordinary ways of speaking. Ordinarily, we talk of outer actions that are rational or irrational. For instance, it is two-boxing that is said to be rational or irrational rather than (a tokening of) some intentional mental state with that outer action as content. Moreover, in any textbook on decision theory, it's outer actions rather than intentional mental states which populate the option columns in decision matrices.

Second, **Subjective Decisions** doesn't fit well with a popular picture of rational deliberation. According to this picture, defended in Skyrms (1990), Arntzenius (2008), and Joyce (2012), the agent's credences for options – conceived of as outer actions – change in response to decision theoretic evaluations of those options. Instead, on **Subjective Decisions**, it appears that rational deliberation consists in the agent making a *decision* that is evaluated best; subsequently becoming confident of executing that decision; then re-evaluating her options with her updated credences, and again making a decision that is evaluated best, and so on. Importantly, only on the latter picture is there *decision instability*. I will talk more about this consequence in Chapter 5.

Third, **Subjective Decisions** has consequences on the demandingness of rational obligation. To demonstrate this, consider the following two cases. First, suppose that Harry is deliberating about which out of two boxes – box 1 and box 2 – to open. Suppose that he is allowed to open only one. Harry knows that box 1 contains £1,000 and that box 2 contains £5,000. Suppose Harry decides on box 2 but owing to his forgetfulness (or perhaps he gets distracted) he opens box 1. The second case is the same as the first case except that the agent, Sally, decides to open box 2, but because she is under tremendous stress she mistakenly opens box 1. It's tempting to say that Harry straightforwardly fails to do as he ought. It's also tempting to say that Sally fails to do as she ought but is excused due to extenuating circumstances, where this implies that she is not the appropriate target of blame. However, if options are decisions, then we can't say either of these things. Instead, we must say that Harry and Sally have done as they ought, for they have both decided to open box 2. If options are decisions, then it turns out that rational obligation is less demanding than you might have thought.

Fourth, conceiving of options as decisions has consequences regarding what Hedden (2015) calls *Diachronic Tragedy*. In the phenomenon of Diachronic Tragedy, an agent performs a sequence of actions which is worse (by her lights) than an alternate sequence of actions that she could have performed. For instance, consider the following case:

> *The Money Pump*. You have intransitive preferences: you prefer Apple Pie to Blueberry Pie, Blueberry to Cherry Pie, and Cherry Pie to Apple Pie. You

> start off with an Apple Pie. You will be offered the following three deals in succession (assuming you accept each earlier deal):
>
> > Deal 1: receive a Cherry Pie in exchange for your Apple Pie and 10 cents.
> >
> > Deal 2: receive an Blueberry Pie in exchange for your Cherry Pie and 10 cents.
> >
> > Deal 3: receive an Apple Pie in exchange for your Blueberry Pie and 10 cents.

If you act on your preferences, then you accept the first deal because you prefer a Cherry Pie to an Apple Pie. However, for analogous reasons, you accept all three deals, which means you end up with what you started with – an Apple Pie – despite your outlay of 30 cents. You have performed a sequence of actions that is worse (by your own lights) than an alternate sequence of actions you could have performed.[53] In response to this case, it's tempting to say that intransitive preferences are irrational precisely in virtue of leading to Diachronic Tragedy. But what exactly is wrong with Diachronic Tragedy? In turn, it's tempting to say that the sequence of actions (which consists of accepting all three deals) is practically irrational by decision theory's lights. However, Hedden (2015) argues that if options are decisions, then this is unavailable because it is *decisions* that are (ir)rational, *not* sequences of actions. (Of course, there might be other ways of saying that there's something wrong with Diachronic Tragedy. But the point is that *qua* theory of rational *ought* there is nothing wrong with it.)

So **Subjective Decisions** has consequences for our ordinary ways of talking about rational obligation, for the correct picture of rational deliberation, for the demandingness of rational obligation, and for the phenomenon of Diachronic Tragedy.[54] In Chapter 5 I examine **Subjective Decisions** and the correct picture of

[53] See Davidson et al (1955) for the original money pump.

[54] Another consequence, implied by my objection to diachronic rivals in section 2, is that **Subjective Decisions** entails that predicted future irrationality is always relevant. It *doesn't* entail, for instance, that an agent's options are extended actions whose decision theoretic

rational deliberation in more detail. Before that, in the next chapter, I deal with the sort of case that I put aside in Chapter 2, section 6 – that is, cases where the agent is uncertain about being able to make a decision for some outer action.

evaluations would ignore the possibility that the agent irrationally starts but fails to complete the extended action. If options are decisions, then *that* possibility will always be factored into the EU's of the decisions.

Chapter 4 – The Missing Option Puzzle

1. Introduction

Suppose Brenda is deliberating about whether to ford a creek. She can in fact ford the creek, but she is uncertain that she can even *decide* to ford a creek. Perhaps she thinks an evil demon might strike her down just as she is about to decide on it. It seems Brenda should have a ford-the-creek-like option, but what is it exactly? There appears to be no suitable ford-the-creek-like option. You might think that the ford-the-creek-like option is *decide to ford the creek*. This is a mistake because decision theory evaluates it in a way that wrongly ignores Brenda's doubts about being able to decide to ford the creek. There appears to be a missing option. This is the Missing Option Puzzle. A similar puzzle is discussed by Pollock (2002).

In Chapter 2 (section 6) I put aside Brenda-type cases, that is, I put aside cases where the agent assigns some (but not all) of her credence to being unable to make a certain decision. I did this in response to an objection to what I was then proposing, namely, **Subjective Actions** along with the sophisticated formulation of decision theory. So my conclusions from Chapter 2, and the subsequent refinement in Chapter 3, are made under the assumption that for every decision, the agent is either certain she can make it or certain she can't make it. Now I want to see what an agent's options are when this assumption fails. This involves trying to resolve the Missing Option Puzzle. Here I am starting from a blank slate again. I won't assume a particular account of options. And I will revert back to the naïve formulation of decision theory, according to which, an agent ought to realise an option that is best.

In response to the Missing Option Puzzle, I will defend what I call the *Counterfactual Strategy*, which says that the ford-the-creek-like option is the counterfactual *if Brenda were able to decide to ford the creek, then she would decide to ford the creek*. I will draw out a number of interesting consequences of this view, in particular, that the Certainty Constraint does not hold for Brenda-type cases. I will also combine the Counterfactual Strategy with **Subjective Decisions** to offer a comprehensive account of options, applicable even to Brenda-type cases. Finally, I'll use the Counterfactual Strategy to pre-empt an objection to the previous chapters.

In section 2, I outline the sort of decision theory I'll be assuming for this chapter (a simple EDT). In section 3, I present the Missing Option Puzzle. In section 4, I work through four natural but mistaken responses to the puzzle. The first response rejects my assumption that there is a ford-the-creek-like option. The second response dismisses the puzzle case as far-fetched. The third response says that Brenda should be modelled as being uncertain about which decision problem she faces. The fourth response says that the ford-the-creek-like option is *Brenda fords the creek or gets struck down*. I'll argue that all these responses fail. In section 5, I present my solution: the Counterfactual Strategy. Then I turn to objections. In section 6, I defend my solution against the objection that says there is a revenge puzzle generated by a case where the demon might prevent the truth of the counterfactual. In section 7, I defend my solution against the objection that says it has the following odd consequence: for some cases, decision theory ends up recommending a counterfactual. That completes my discussion of the Missing Option Puzzle. Finally, I'll discuss consequences for the previous chapters' discussion. There is also an appendix outlining Pollock's solution to the puzzle and why I think it fails.

2. Evidential Decision Theory

It'll be useful to have a specific version of decision theory in front of us. As I'll understand it here, decision theory assumes that the agent has *credence* and *utility* functions defined in the first instance over single possible worlds. Each world w has a credence which represents the agent's degree of belief in that world being the actual world, Cr(w). These credences fall on a scale between zero and one, and they sum to one. Each world w also has a utility which measures the agent's desire for that world to be the actual world, U(w). These utilities fall on a linear scale with arbitrary zero and unit.

A few bits of terminology will be useful in what follows. First, a set of worlds is a *proposition*. The proposition that grass is green is the set of worlds where grass is green. Credence for a proposition is simply the sum of the credences for each world in that set. Second, for propositions P and Q, the agent's *conditional credence P given Q*, Cr(P|Q), is defined in the normal way as Cr(P&Q)/Cr(Q) when Cr(Q)>0. Third, the (evidential) expected utility, EU, of a proposition P is an average of the utilities of worlds, weighted by the conditional credence in each world given P.[55] Formally-speaking, the EU of a proposition P is as follows:

$$\sum_{w} Cr(w|P)U(w)$$

Decision theory has three components. First, a specification of which propositions are the agent's options. Second, an evaluation of options in terms of their EU's. The third component is a prescription to realise an option that is evaluated best. That is, decision theory says that an agent ought to realise the option with the greatest EU. More precisely, given that options are *propositions* in this framework, the agent ought to *make true* the option-proposition that is evaluated best. This propositional construal of decision theory's prescription will be important when I come to consider objections to the Counterfactual Strategy.[56]

3. The Missing Option Puzzle

The Missing Option Puzzle is the challenge of saying what the agent's options are in a case where she is uncertain that she can decide on a particular action. For instance, consider the following case (which is a fleshed out version of the Brenda case offered at the start of this chapter):

[55] Strictly-speaking, given that the definition of conditional credence is defined over *propositions*, the weights are conditional credences in the *singleton set* of each world given P. But I won't be careful about this difference.

[56] I'm assuming a simple Evidential Decision Theory from Lewis (1981). For the classic presentation of Evidential Decision Theory, see Jeffrey (1983).

> **Brenda's case**. One day, Brenda hikes from her home and into the countryside. After a while, she comes to a raging creek which blocks her path. Brenda would like to get to the other side of the creek so that she can continue her hike. She slightly prefers this to returning home and spending the rest of the day watching TV. Brenda is certain that she can go home. However, she is 50-50 on whether she is able to decide to ford the creek. That's because she thinks that Black might strike her down just as she is about to decide to ford the creek, thus preventing her from deciding to ford the creek – a very bad outcome by Brenda's lights. In actual fact, Brenda can both go home and ford the creek.[57]

One reaction to Brenda's case is to say that it is irrelevant because far-fetched. However, note that Brenda *can in fact* decide to ford the creek – it's just that she *thinks* she *might* be unable to decide to ford the creek. Although a case where Brenda will get struck down may be far-fetched, surely this doesn't apply to a case where Brenda *thinks* she *might* get struck down. I'll have more to say about this in section 4.

I take it that the intuition in Brenda's case is that she ought to go home. For fording the creek carries a huge risk from her perspective – she might get struck down – and this risk isn't worth taking in order to continue the hike. How do we apply the decision theoretic framework so that it delivers a sensible verdict? In particular, what are the options? Clearly, *go home* is one option. It also looks like there's an option corresponding to ford the creek. But what exactly is this ford-the-creek-like option?

Naively, one might think that the ford-the-creek-like option is *decide to ford the creek*. However, the problem with this suggestion is that decision theory misevaluates it. That's because its EU is a weighted average of the utilities of worlds – *in Brenda's doxastic space* – where she decides to ford the creek. But this completely ignores the worlds where Brenda gets struck down, because there is no world in Brenda's

[57] This is a development of a case in Hedden (2012), which in turn is a development of a Frankfurt case (see Frankfurt 1969).

doxastic space such that Brenda decides to ford the creek and gets struck down – after all, Brenda thinks that if she decides to ford the creek, then she's somehow *avoided* getting struck down by Black. The problem is that the evaluation of *decide to ford the creek* ignores Brenda's very relevant doubts about being able to decide to ford the creek. (Note that *ford the creek* won't do either. For *a fortiori* the set of worlds in Brenda's doxastic space where she fords the creek excludes worlds where she gets struck down.)

To see that this misevaluation of *decide to ford the creek* leads to trouble, note that its EU reduces to the utility of a world where Brenda continues the hike – because in every world in her doxastic space in which she decides to ford the creek, she also continues the hike. Note also that the EU of *go home* reduces to the utility of a world where Brenda watches TV – because in every world in her doxastic space where she goes home, she also watches TV. Given that *ex hypothesi* the utility of a world where Brenda continues the hike is greater than the utility of a world where she watches TV, this means that decision theory says that Brenda ought to decide to ford the creek – the wrong verdict. In sum, the ford-the-creek-like option can't be *decide to ford the creek* because decision theory wrongly evaluates it, with the upshot that it delivers the wrong verdict in Brenda's case.

The Missing Option Puzzle is the challenge of specifying the ford-the-creek-like option in Brenda's case such that decision theory delivers a sensible verdict. More generally, the puzzle concerns a case where the agent doubts that she can decide on an action A because X might befall her. The puzzle is the challenge of finding the A-like option that is evaluated correctly by decision theory, namely, in a way that takes into account the agent's doubts about being able to decide on A. The candidate *decide to A* isn't up to the job because it's false at doxastically possible worlds where X befalls the agent, so such worlds are deemed irrelevant when evaluating it – this seems to completely ignore the agent's very relevant doubts about being able to decide on A. There appears to be a missing option and the challenge is to find it.

To justify the intuition that Brenda ought to go home, I said that fording the creek carries a huge risk from Brenda's perspective – she might get struck down – and this risk isn't worth taking in order to continue the hike. Let me try another way to justify the intuition. Recall that decision theory is a theory of what an agent ought to do under conditions of uncertainty. So just as it's relevant what the agent thinks ford the creek might lead to, it's surely relevant that the agent doubts that she is able to decide to ford the creek – treating these two sorts of uncertainty differently appears arbitrary. In other words, everyone can agree, in the decision theoretic framework, that a supernatural being who might strike down Brenda *in response to* her fording the creek is relevant. Clearly, it's a bad-making feature. Given this, if it's *not* relevant that Black might strike down Brenda just as she's about to decide to ford the creek, then things seem very arbitrary. A being who might strike down Brenda in *response* to her fording the creek is *relevant*, whereas a being who might intervene *before* Brenda fords the creek (just before she decides) is *not* relevant. That's a strange picture of what counts as relevant.

You might be thinking that there is some other species of mental action, other than a *deciding*, that is a suitable ford-the-creek-like option in Brenda's case. This may be right, but it is ultimately irrelevant. For instance, perhaps if Brenda fords the creek, then she *wills* to ford the creek which in turn would lead to her *deciding* to ford the creek which in turn would lead to her actually *fording the creek*. As I've set up the case, Black might intervene just before a decision to ford the creek, which is plausibly interpreted as intervening *after* willing to ford the creek but *before* deciding to ford the creek. Given this, you might think that the ford-the-creek-like option in Brenda's case is *will to ford the creek*. It would seem that Brenda assigns credence to getting struck down given that she wills to ford the creek, so decision theory's evaluation of *will to ford the creek* correctly takes into account Brenda's doubts about being able to decide to ford the creek. However, it's obvious that this merely forces a tweak to the puzzle – simply imagine that the agent doubts that she is able to will to ford the creek, because she thinks Black might strike her down just before she wills to ford the creek. Then *will to ford the creek* can't be the ford-the-creek-like option, because decision theory evaluates it by looking only at the worlds where Brenda wills to ford the creek, which excludes worlds where Brenda gets struck down. This is tantamount to ignoring Brenda's doubts about being able to will to ford the creek, which are intuitively relevant. The same goes for any other mental action one might propose: trying to ford the creek, intending to ford the creek etc. In other words, if another mental action is a suitable ford-the-creek-like option in Brenda's case, then it generates a revenge puzzle – simply consider a case where the agent doubts that she can perform the mental action in question because Black might strike her down before she does it.

I've presented the puzzle for a particular version of decision theory – namely, Evidential Decision Theory. But the puzzle extends beyond this.

First, decision theory is an example of a *consequentialist theory*, that is, a theory that evaluates an option by looking at the *value* of worlds that receive a positive conditional *probability* given the option, for some species of *value* and *probability*. The Missing Option Puzzle applies to all consequentialist theories. Simply imagine a case where there is some probability assigned to the agent being unable to decide on some action A. Perhaps Black might strike her down if she is about to decide on A. What is the A-like option? It can't be *decide to A* because it would be evaluated by looking only at worlds – that is, worlds that are assigned nonzero probability – where the agent decides to A, which excludes worlds where she is struck down. This is tantamount to ignoring the probability assigned to the agent being unable to decide to A, which is intuitively relevant. In what follows I will focus only on the Missing Option Puzzle *as applied to decision theory*. That's because I think the puzzle hits decision theory harder than other consequentialist theories. And that's because the case in question is clearly not far-fetched when the probability in question is credence. In contrast, when the probability in question is some more objective species of probability, then I think it is more debatable whether the relevant case is far-fetched. So the puzzle is more pressing for the decision theorist.

Second, the Missing Option Puzzle affects Causal Decision Theory as well as Evidential Decision Theory. As I'll understand Causal Decision Theory, it says that an agent ought to realise an option with the best causal EU, where the causal EU of a proposition P is as follows:

$$\sum_w Cr(P \rightarrow w)U(w)$$

Here, '→' refers to the counterfactual conditional.[58] Now, if Brenda's ford-the-creek-like option is *decide to ford the creek*, then Causal Decision Theory misevaluates it. For Brenda assigns credence 0 to the proposition *if Brenda were to decide to ford the creek, then she would get struck down*. That's because she thinks that if she decides to ford the creek, then she's somehow avoided getting struck down. So the evaluation of *decide to ford the creek* ignores the utilities of worlds where Brenda gets struck down – the utilities of such worlds are weighted by 0 in the calculation of its causal EU. Hence Brenda's doubts about being able to decide to ford the creek are ignored by an evaluation of *decide to ford the creek*. The fundamental problem is that a world where Brenda gets struck down does not count as a *possible outcome* of deciding to ford the creek – on any construal of "possible outcome", evidential or causal – and decision theory evaluates an option by looking only at its possible outcomes. This results in decision theory – in an evidential or causal form – misevaluating *decide to ford the creek*. In what follows, I'll assume the simple Evidential Decision Theory outlined in section 2 for the sake of having a concrete version of decision theory in front of us. I don't think anything substantial hinges on this.

4. Possible Responses

Before getting to my solution to the Missing Option Puzzle, let me go through four other responses. The first rejects my assumption that Brenda has a ford-the-creek-like option; the second dismisses the puzzle case as far-fetched; the third models Brenda as being uncertain about which decision problem she faces; the fourth involves conceiving of Brenda's ford-the-creek-like option as the proposition *that Brenda fords the creek or gets struck down*. I'll argue that all four responses fail.

[58] This Causal Decision Theory appears in Gibbard & Harper (1978). For a survey of causal decision theories, see Joyce (1999).

The first response questions whether there really is a puzzle here. After all, the puzzle is generated by assuming that Brenda has a ford-the-creek-like option and noting that the obvious candidates (ford the creek, decide to ford the creek) are all misevaluated such that decision theory delivers the verdict that she ought to do the ford-the-creek-like thing – contrary to the intuition that she ought to go home. However, if we drop the assumption that there is a ford-the-creek-like option, then it's easy to guarantee that decision theory delivers the verdict that Brenda ought to go home – for then it seems that Brenda will have only one option, namely, *go home*!

To demonstrate why there has to be a ford-the-creek-like option, consider a case exactly like Brenda's except that the agent – call her Cathy – *very strongly* desires to continue her hike. Consider her desire strong enough so that the risk of getting struck down by Black is worth it. Then I think it's intuitive that Cathy ought to realise the ford-the-creek-like option. If decision theory is to deliver this verdict, then there needs to be a ford-the-creek-like option in Cathy's case. So there is a ford-the-creek-like option in Cathy's case. Now if Cathy has a ford-the-creek-like option, then so does Brenda, because all that's different between these agents is the desirability of continuing the hike. This sort of difference – a difference in the desirability of a possible outcome – shouldn't make a difference to the agent's options. So Brenda also has a ford-the-creek-like option.

Now for the second response. You might say that there is something wrong with Brenda's case. The most obvious way of pushing this is to say that Brenda's case is too far-fetched to warrant serious consideration. However, as I've already said, the relevant case is not far-fetched. For Brenda assigns 0.5 credence to Black striking her down. A case where Brenda is struck down by Black may be unrealistic; but surely a case where Brenda *thinks* she *might* get struck down isn't far-fetched.

Moreover, there are two ways to make the case even less far-fetched. First, I appeal to a supernatural being – Black – with the ability to prevent Brenda from deciding to ford the creek. There are other things that might be able play this role. For instance, pathological psychological phenomena such as phobias, addictions and indecisiveness. Suppose that Brenda is hydrophobic. Then she might be unable to bring herself to genuinely decide to ford the creek. After all, fording the creek involves getting *in* the creek.[59] Also, the agent's own beliefs in the inefficacy of her decisions might also play the role of Black. Suppose Brenda believes her decision to ford the creek would be inefficacious. Then she might be unable to decide to ford the creek. A final example: if instead of contemplating whether to ford the creek, Brenda is contemplating something she finds morally grotesque e.g. torturing kittens, then she might be unable to genuinely decide to do this.

A second way of making the case less far-fetched is as follows. There's nothing special about it being 0.5 that Brenda assigns to being unable to decide to ford the creek – it can be much smaller. For instance, suppose Brenda assigns just 0.05 credence to getting struck down by Black. Presumably, we often assign such credences to far-out possibilities like Black's intervention. A credence of 0.05 will do just as well as a credence of 0.5, because however much credence is assigned to being unable to decide to ford the creek, there needs to be a ford-the-creek-like option whose EU is sensitive to such doubts, otherwise decision theory misevaluates it.

Another way to push the objection that Brenda's case is somehow illegitimate is to say that, though not far-fetched, it's simply not the sort of case to which decision theory applies. You might think that decision theory applies to cases where there is some set of actions the agent knows she can realise, and this isn't so in Brenda's case. The problem with this response is that it severely limits the applicability of decision theory if, as conceded above, the cases are not far-fetched. Admittedly, even if not far-fetched, Brenda's case is *nonstandard* in a certain sense – in the sense that the agent doubts that she can make a decision. In applying decision theory to such nonstandard cases, this chapter can be seen as following in the

[59] Pollock (2002) makes the same point in his version of the puzzle.

tradition of a substantial body of work that tries to apply decision theory to nonstandard cases.[60]

A third response to the puzzle rejects an assumption that has been implicit thus far. The assumption is that there is only one option set in Brenda's case. Instead, you might think that Brenda's case should be modelled such that Brenda is uncertain about which decision problem she faces, because she is uncertain if Black is lurking. If Black is lurking (i.e. is ready to pounce and strike Brenda down just before she decides to ford the creek), then Brenda is facing a decision problem where the only option is *go home*. If Black isn't lurking, then Brenda is facing a decision problem where the options are *go home* and *ford the creek*. (Alternatively, one might frame this in terms of decisions, but I'll stick with nonmental actions here – nothing hinges on this.) To generate prescriptions for Brenda, here's what we do. First, conditionalize (Brenda's credence function) on Black not lurking. Then apply decision theory in the normal way with the options as *ford the creek* and *go home*. Decision theory will say that Brenda ought to ford the creek. That's the right prescription because the prescription is relative to the decision problem where there is no Black. Second, conditionalize on Black lurking. Then apply decision theory in the normal way with the single option as *go home*. Obviously, decision theory will say that Brenda ought to go home. That's the right prescription because the prescription is relative to the decision problem where Black is lurking. So this approach gives us two prescriptions – each relative to a decision problem, and Brenda is uncertain about which decision problem she's in.

Does this strategy work? I'm happy to set decision theory's aim lower, as this approach does. On this approach, decision theory doesn't deliver a verdict about what Brenda ought to do. It merely delivers a verdict about what she ought to do *given which decision problem she faces*, about which she might be uncertain. Although I'm happy to do this, I want a theory that delivers a definitive verdict about what Brenda

[60] See Elga (2010) and Hare (2010) for discussion of applying decision theory in the absence of precise probabilities and utilities respectively. See Weirich (2004) and Bradley (2017) for attempts to apply decision theory to agents who are limited (in time and intelligence).

ought to do. So the question is: can decision theory, on this approach, be part of a *larger* theory that delivers a sensible verdict in Brenda's case?

I don't think it can. The fundamental problem is that nowhere has the disutility of Brenda's getting struck down been taken in to account. In each decision problem the evaluation of each option completely ignores this. To see this, consider what a larger theory would look like. Presumably its verdict would be a function of two things, first, the EU's of each option in each decision problem, and second, Brenda's uncertainty about which decision problem she's in. Let a *candidate* be something that's an option according to at least one decision problem that the agent thinks she might be facing. Presumably the theory says something like the following: Brenda ought to realise the candidate with best expected expected-utility (EEU), where the EEU of a candidate is an average of the EU's of the candidate in each decision problem D in which it occurs as an option, weighted by the agent's credence in facing D. This larger theory delivers the wrong verdict in Brenda's case. For imagine we flesh out Brenda's case so that a world in which Brenda continues the hike has a utility of 10, a world in which Brenda gets struck down has a utility of -50, and a world in which she watches TV has a value of 0. This reflects the fact that Brenda slightly prefers continuing the hike to watching TV, but she would really hate getting struck down by Black. Given these utility assignments, the larger theory delivers the wrong verdict. Let w be a world where Brenda continues the hike; let B be the proposition that Black is lurking. Then the EEU of the proposition that Brenda fords the creek is:

$$U(w)*Cr(\text{not-}B) = 10*0.5 = 5$$

Why? The proposition that Brenda fords the creek occurs as an option in only one decision problem – the one where there is no Black – so its EEU is its EU in that decision problem weighted by Brenda's credence for being in that decision problem. It's EU in that decision problem is just the utility of a world where Brenda continues the hike, and her credence for being in that decision problem is just her credence that there is no Black, hence the above.

In comparison, consider the EEU of the proposition that Brenda goes home. Let w' be a world where Brenda watches TV. Then the EEU of *go home* is:

U(w')*Cr(not-B) + U(w')*Cr(B) = 0*0.5 + 0*0.5 = 0

Why? The proposition that Brenda goes home appears as an option in both decision problems. Its EEU is its EU in each decision problem weighted by Brenda's credence for being in that decision problem. It's EU in each decision problem is simply the utility of a world where Brenda watches TV, and her credence for being in each decision problem is just her credence that there is or isn't Black lurking. Hence the above. So the larger theory says that Brenda ought to ford the creek – the wrong verdict.

What's happened here is that nowhere has the disutility of Brenda getting struck down by Black been taken into account. So it's no surprise that the larger theory wrongly says that Brenda ought to ford the creek. This demonstrates something important about an adequate solution. It's not just that we want to take into account Brenda's doubts about being able to decide to ford the creek. This is done by the above response but it's not an adequate response to the puzzle. More precisely, we want to take into account Brenda's doubts about being able to decide to ford the creek *in the right way*, in particular, in a way that takes into account the very poor utility assigned to worlds where she gets struck down. It's because this second response takes no account of this that it is inadequate.

The fourth response says that Brenda's ford-the-creek-like option is the disjunctive proposition *Brenda fords the creek or gets struck down*. More generally, this response says that if the agent doubts that she can decide on an action A, because she thinks X might befall her before she decides on A, then the A-like option is the proposition *Agent A's or X's*. Let's call this response the *Disjunctive Strategy*.

This strategy appears to do well in Brenda's case. For note that Brenda's doxastic space can be partitioned into three regions of interest: first, worlds where Brenda goes home (and watches TV); second, worlds where she fords the creek (and continues the hike); third, worlds where she gets struck down by Black. The proposition *Brenda fords the creek or gets struck down* is true in all and only the worlds in which Brenda fords the creek or gets struck down. This makes its EU an average of the utilities of worlds where she fords the creek or gets struck down, weighted by

Brenda's credence in each world given the disjunction. The disjunction's EU looks appropriately sensitive to Brenda's doubts about being able to decide to ford the creek, because it is sensitive to the utilities of worlds where Brenda gets struck down.

The problem with the Disjunctive Strategy is that, although it may work in Brenda's case, it sometimes delivers options whose EU's are sensitive to the wrong worlds. To see this, consider the following case:

> **Jane's case**. One day, Jane hikes from her home and into the countryside. After a while, she comes to a raging creek which blocks her path. Jane would like to get to the other side of the creek so that she can continue her hike. She slightly prefers this to returning home and spending the rest of the day watching TV. Jane assigns 0.5 credence to being able to decide to ford the creek because she thinks Black might strike her down just as she is about to decide to ford the creek. In actual fact, Jane can both go home and ford the creek. So far this is the same as Brenda's case. However, unlike Brenda's case, Jane is also unsure that she can decide to go home. In particular, she assigns 0.8 credence to being able to decide to go home because she thinks that Black might strike her down just as she is about to decide to go home. As before, getting struck down by Black is a very bad outcome by the agent's lights.

It looks like there should be a ford-the-creek-like option and a go-home-like option. However, for the same reason that *decide to ford the creek* won't do for the ford-the-creek-like option, *decide to go home* won't do for the go-home-like option. So we have two missing options here. What does the Disjunctive Strategy say? Recall that it says that if the agent doubts that she can decide on an action A, because she thinks X might befall her, then the A-like option is the proposition *Agent A's or X's*. So the Disjunctive Strategy says that the two options in this case are the following propositions:

> Jane goes home or gets struck down
>
> Jane fords the creek or gets struck down

Note that Jane's doxastic space can be partitioned into four regions of interest: first, worlds where she goes home (and watches TV); second, worlds where she fords the creek (and continues the hike); third, worlds where she gets struck down by Black just as she is about to decide to go home; fourth, worlds where she gets struck down by Black just as she is about to decide to ford the creek. The problem is that both of the options are true at the third and fourth regions. The result is that each option's EU is sensitive to the utilities of worlds it shouldn't be sensitive to. The EU of the ford-the-creek-like option is sensitive to the utilities of worlds in which Jane is about to decide to go home but gets struck down – this isn't right. The EU of the go-home-like option is sensitive to the utilities of worlds in which Jane is about to decide to ford the creek but gets struck down – this also isn't right. So the Disjunctive Strategy, though it works in Brenda's case, delivers the wrong options in Jane's case, and for that reason, it's not an adequate response to the puzzle.

(In response to this objection to the Disjunctive Strategy, you might think that the relevant disjunctions simply need to be reworked. Instead of *Jane fords the creek or gets struck down*, you might think the relevant disjunction is *Jane fords the creek or gets-struck-down-just-as-she-is-about-to-decide-to-ford-the-creek*. Similarly for the go-home-like option. I am sympathetic to this response. However, more needs to be said about what it is to be "struck down just as you are about to decide to ford the creek" because it's not immediately obvious what this amounts to. It's tempting to say that it consists in the fact that you would have decided to ford creek had nothing intervened. But then this reworking of the Disjunctive Strategy is effectively the Counterfactual Strategy, which I propose in the next section.)

5. The Counterfactual Strategy

I propose that Brenda's ford-the-creek-like option is the counterfactual proposition *if Brenda were able to decide to ford the creek, then she would decide to ford the creek*. More generally, I propose that whenever an agent doubts that she can decide on an action

A, then the A-like option is the counterfactual *if the agent were able to decide to A, then she would decide to A*.

Recall that Brenda's doxastic space can be partitioned into three regions of interest: first, worlds where Brenda goes home (and watches TV); second, worlds where she fords the creek (and continues the hike); third, worlds where she gets struck down by Black. The counterfactual proposition *if Brenda were able to decide to ford the creek, then she would decide to ford the creek* is trivially true in worlds where Brenda fords the creek (because in such worlds she decides to ford the creek). It is also true in worlds where she gets struck down. This is because, in such a world, the nearest world such that Brenda is able to decide to ford the creek, i.e. where there is no demon ready to strike her down, she manages to decide to ford the creek. Moreover, the counterfactual is false in worlds where Brenda goes home. This is because: at a world where Brenda goes home, the closest world where Brenda is able to decide to ford the creek is still a world where Brenda goes home. So the proposition is true in all and only the worlds in which Brenda either continues the hike or gets struck down. This means that the proposition's EU will be an average of the utilities of the worlds where Brenda continues the hike or gets struck down, weighted by her credence in each world given the counterfactual. I submit that this adequately takes into account Brenda's doubts about being able to ford the creek.

To justify this latter claim, consider the prescription the Counterfactual Strategy makes decision theory deliver in Brenda's case. Let's flesh out Brenda's case so that the world in which Brenda continues the hike has a utility of 10, the world in which Brenda gets struck down has a utility of -50, and the world in which she watches TV has a value of 0. This reflects the fact that Brenda slightly prefers continuing the hike to watching TV, but she would really hate getting struck down by Black. Given this, the EU of the proposition that Brenda goes home reduces to the utility of a world where she watches TV, so it's 0. Now let's work out the EU of the counterfactual. Its EU is an average of the utilities of worlds where Brenda either fords the creek or gets struck down, weighted by the agent's credence in each world given the counterfactual. Let F stand for the proposition that Brenda fords the creek; let S stand for the proposition that Brenda gets struck down; let C stand for the counterfactual; let w^F and w^S stand for a world in which Brenda fords the creek and a world in which Brenda gets struck down respectively. Then the EU of the counterfactual is:

$$Cr(F|C)EU(F\&C) + Cr(S|C)EU(S\&C)$$
$$= Cr(F|C)(10) + Cr(S|C)(-50)$$

The first line represents the EU of the counterfactual by definition of EU in section 2, and by the fact that F and S form a partition over C (this is an instance of Lewis' *Rule of Averaging*, Lewis 1981, pp.6-7). The EU of F&C is 10 because every world in F&C has a utility of 10 – this is how I fleshed out the case immediately above Similarly for S&C and -50. At this stage, we can't put a precise figure on the EU of the counterfactual because we don't know the relevant conditional probabilities. This is the case even though we know that Brenda is 50-50 on whether she is able to decide to ford the creek. What we do know is that if Brenda's credence assignment to getting struck down is *equal to* 1/5 of her credence assignment to fording the creek, then the EU of the counterfactual will be 0, in which case decision theory says either option is permissible. (We know this because the ratio between Brenda's unconditional credence in getting struck down and Brenda's unconditional credence in fording the creek is the same as the ratio between Cr(S|C) and Cr(F|C) – that's because the counterfactual is true in all and only the worlds in which Brenda gets struck down or fords the creek.) Alternately, if Brenda's credence assignment to getting struck down is *more than* 1/5 of her credence assignment to fording the creek, then the EU of the counterfactual will be negative, in which case decision theory says that Brenda ought to go home. If Brenda's credence assignment to getting struck down is *less than* 1/5 of her credence assignment to fording the creek, then the EU of the counterfactual will be positive, in which case decision theory says that Brenda ought to choose the ford-the-creek-like option. This, I think, delivers sensible verdicts: Brenda ought to go home unless she assigns much less credence to getting struck down than to fording the creek.

The Counterfactual Strategy says that whenever an agent doubts that she can decide on an action A, then the A-like option is the counterfactual *if the agent were able to* *decide to A, then she would decide to A*. This gets Brenda's case right, but does it get Jane's case right? This is the case the Disjunctive Strategy faltered on. Recall that Jane's doxastic space can be partitioned into four regions of interest: first, worlds where she goes home (and watches TV); second, worlds where she fords the creek (and continues her hike); third, worlds where she gets struck down by Black just as she is about to decide to go home; fourth, worlds where she gets struck down by Black just as she is about to decide to ford the creek. The Disjunctive Strategy says the options are the following propositions:

> Jane goes home or gets struck down
>
> Jane fords the creek or gets struck down

The problem is that both of these propositions are true in the third and fourth regions i.e. the regions where she gets struck down just as she is about decide to go home and ford the creek respectively. The result is that each option's EU is sensitive to the utilities of worlds it shouldn't be sensitive to. In contrast, the Counterfactual Strategy says the options are the following propositions:

> If Jane were able to decide to go home, then she would decide to go home
>
> If Jane were able to decide to ford the creek, then she would decide to ford the creek

These propositions are true at the right regions. The first is trivially true in worlds where Jane goes home (because in such worlds she decides to go home). It is also true in worlds where Jane gets struck down just as she is about to decide to go home. That's because it's true that, in the closest world in which Jane can go home, so where there is no demon, she manages to decide to go home. Moreover, it is true only in these worlds. First, it is not true in any world in which Jane fords the creek. Intuitively, in that world it's not the case that the closest world in which Jane can go home is one in which she decide to go home – for in that world she still fords the creek. Second, it's not true in any world in which she gets struck down just as she is about to decide to ford the creek. Intuitively, in that world it's not the case that in the closest world in which Jane can decide to go home is one in which she decides to go home – for in that world she still gets struck down just as she is about to decide to ford the creek. For parallel reasons, the proposition *if Jane were able to decide to ford the creek, then she would decide to ford the creek* is true in all and only worlds where Jane either fords the creek or gets struck down just as she is about to decide to ford the creek. So the Counterfactual Strategy delivers options in Jane's case which are true at the correct regions in Jane's doxastic space and which consequently get evaluated correctly.

Let me be a bit more precise about the relevant counterfactual. I say the ford-the-creek-like option (in Brenda's case) is the counterfactual proposition (1) *if Brenda were able to decide to ford the creek, then she would decide to ford the creek.*[61] The counterfactual proposition (1) is to be distinguished from the following propositions:

(2) Brenda makes true the proposition that *if she were able to decide to ford the creek, then she would decide to ford the creek*;

(3) Brenda has the dispositional property associated with the counterfactual *if she were able to decide to ford the creek, then she would decide to ford the creek.*

((2) is a proposition about a proposition. (3) is a tensed proposition – true at worlds *and times* where *and when* Brenda has the relevant dispositional property, and false at others.) To see the difference between these three propositions, consider a world w where Brenda is just about to decide to ford the creek but she gets struck down by Black, where this is understood as instant death. In w, (1) is true – that is, if Brenda were able to decide to ford the creek, then she would decide to ford the creek. That's because: in the closest world in which Brenda is able to decide to ford the creek, Brenda decides to ford the creek.

[61] There are suppressed time-indices here. More precisely, I say the ford-the-creek-like option is the following counterfactual:

> If Brenda were able-at-t_c to decide-at-t_i-to-ford-the-creek, then she would decide-at-t_i-to-ford-the-creek.

The time t_c is the time at which Brenda faces the choice (i.e. the time at which Brenda comes to a raging creek and starts to deliberate about what to do); t_i is the time immediately after t_c.

However, (2) is false in w – that is, Brenda doesn't make true the proposition that if she were able to decide to ford the creek, then she would decide to ford the creek. That's because: making true a proposition P is acting in such a way that P is true; when Brenda gets struck down, she doesn't act at all; so in w she doesn't make true the counterfactual proposition.

Moreover, (3) is true in w at some times and false at others. Let the time Black strikes Brenda down be t_x. Then Brenda has the dispositional property before t_x but lacks it after t_x. Before being struck down, Brenda has the dispositional property – that is, she is such that if Black were not to intervene she would go on to decide to ford the creek. After being struck down, Brenda doesn't have the dispositional property, because she's not alive.

So (1), (2), and (3) are distinct propositions. I'm saying that Brenda's ford-the-creek-like option is (1) i.e. if Brenda were able to decide to ford the creek, then she would decide to ford the creek. In the next two sections, I turn to possible objections.

6. Objection One: The Revenge Puzzle

In section 3 I said that *will to ford the creek* might be a ford-the-creek-like option which is evaluated correctly in Brenda's case. But I said that this doesn't constitute a solution to the puzzle because it's susceptible to a revenge puzzle: simply imagine a case where Brenda thinks an evil demon might strike her down just before she *wills* to ford the creek. Is the Counterfactual Strategy susceptible to a similar revenge puzzle?

There is no revenge puzzle for the Counterfactual Strategy because the counterfactual doesn't describe an event in the process that results in Brenda's fording the creek. Let me flesh this out. When Brenda fords the creek, inner events (e.g. neurological events, a decision, etc.) lead to outer events (e.g. certain bodily movements, Brenda wading through the water etc.). Let's call this unfolding of inner and outer events the *agential process*. Now consider the dialectic with the proposition *that Brenda wills to ford the creek*. This is evaluated correctly in Brenda's case because it describes an event in the agential process where the event is earlier than Black's possible interruption. However, there is a revenge puzzle for this alleged solution because we can simply imagine a case where Black might strike the agent down *earlier* in the process – *just before* the willing. In such a case, Black's possible interruption is relevant, but an evaluation of an event downstream from the possible interruption completely ignores the interruption. In contrast, the counterfactual proposition does not describe an event in the agential process, so it is not subject to a revenge puzzle: there is no corresponding event in the agential process that we can point to and say "imagine that Black might strike the agent down just before *this* event."

I think this response works, but it leaves some unanswered questions. First, what exactly is this *agential process*? Very roughly, it's the process that ends with fording the creek, and working back, includes trying to ford the creek, deciding to ford the creek, (possibly) willing to ford the creek, and then I'm not sure what else it includes. But there's no need to settle this matter. The counterfactual does not describe an event in this process (whatever this process includes), so there's no possibility of a revenge puzzle of the sort to which *will to ford the creek* is susceptible.

Second, which sorts of possible interruptions (apart from Black striking the agent down) are relevant to an evaluation of the ford-the-creek-like option? Not all interruptions. For instance, Black might persuade the agent to go home (by rational dialogue) just before she is about to decide to ford the creek. But this is not relevant (or at least not obviously relevant) to the evaluation of the ford-the-creek-like option: it is not a bad- or good-making feature of the ford-the-creek-like option. However, there's no need to settle which interruptions are the relevant ones. For striking the agent down *is* relevant, and this is enough to generate the Missing Option Puzzle. Moreover, the counterfactual is not an event in the agential process, so it is not subject to a revenge puzzle – this is so regardless of which interruptions are the relevant ones.

7. Objection Two: Choosing a Counterfactual

You might think that the Counterfactual Strategy has a very odd consequence. As I formulated decision theory, it says that an agent ought to realise an option of maximal EU. In the propositional framework (where options are propositions), this amounts to making true the proposition that's best. So if options are sometimes counterfactual propositions, then agents sometimes ought to make true a counterfactual proposition. For instance, recall Cathy's case, which is exactly like Brenda's case except that she assigns a very high utility to continuing the hike. Given her utility assignment to continuing the hike, it will come out that Cathy ought to make true the counterfactual proposition *if Cathy were able to decide to ford the creek, then she would decide to ford the creek*.

Now Cathy makes that counterfactual true just when she acts in such a way that it is true. So she makes it true just when she decides to ford the creek. So I'm committed to saying that Cathy ought to decide to ford the creek. That seems fine. I stipulated that Brenda can in fact ford the creek. So Cathy can in fact ford the creek as well, which means she can also decide to ford the creek. So the prescription is something that Cathy can fulfil. However, it's natural to wonder what happens when Cathy *can't* decide to ford the creek. Here we have a motivation for the sophisticated formulation of decision theory. On this formulation, decision theory says that an agent ought to do as much as she can of the best option. It ensures that the prescription of decision theory is something that the agent can fulfil (see Chapter 2, section 3). So for Brenda-type cases, I endorse the sophisticated formulation of decision theory. (That this formulation of decision theory is motivated in this setting is not really surprising, because Brenda-type cases always seemed like an objection not to the sophisticated formulation of decision theory but to **Subjective Actions**.) I note, however, that **Subjective Actions**, and its refinement, **Subjective Decisions**, remain accounts of options only for non-Brenda-type cases.

8. Conclusion

At the start of this Chapter, I said that the preceding chapters have assumed that the agent isn't uncertain, of some decision, that she can make it. That is, they have assumed that, for every decision, the agent is either certain she can make it or certain she can't make it. I assumed this in Chapter 2 (section 6) because in a case where the agent *is* uncertain of some decision that she can make it, there is a puzzle for any account of options, namely, the Missing Option Puzzle.

In this chapter I looked at the Missing Option Puzzle. What is the A-like option when the agent doubts that she can decide on A? *Decide to A* can't be the option because decision theory misevaluates it by ignoring the agent's doubts about being able to decide on A. I proposed the Counterfactual Strategy – the A-like option is the counterfactual proposition *if the agent were able to decide to A, then she would decide to A*. This has a number of interesting consequences.

First, if the Counterfactual Strategy is right, then the upshot is that the Certainty Constraint doesn't hold in cases where the agent is uncertain, of some decision, that she can make it. Although it holds in all other cases, in the Brenda-type cases, the Certainty Constraint doesn't hold. This is the upshot because Brenda isn't certain that she can make true the counterfactual. Making true a proposition P is acting in such a way that the proposition is true. So when Brenda decides to ford the creek, she makes true the counterfactual. But when Brenda goes home, she doesn't make true the counterfactual. Also, when Brenda gets struck down, she doesn't make true the counterfactual. That's because: when she gets struck down, she doesn't make *anything* true because she doesn't perform an action. Given that Brenda makes the counterfactual true just when she decides to ford the creek, it follows that Brenda (assuming she isn't conceptually confused) is confident that she can make true the counterfactual to the extent that she is confident that she can decide to ford the creek. Given that Brenda is less-than-certain that she can decide to ford the creek, she is less-than-certain that she can make true the counterfactual. (Nevertheless, I remain committed to the Certainty Constraint in cases where the agent is not uncertain, of some decision, that she can make it.)

Second, the Counterfactual Strategy has something in common with moral luck scepticism. The latter denies that one's degree of moral responsibility is affected by luck. For instance, suppose George kills someone and suppose another agent Georg fails to kill owing to luck. The moral luck sceptic says that both are as responsible as each other. Now this is compatible with both being as responsible as killers normally are, and with neither being responsible at all. However, at least some moral luck sceptics (e.g. Zimmerman 2002), and this is the sort of moral luck sceptic I want to talk about here, say that both are as responsible as killers normally are. Moral luck sceptics of this sort have a hard time finding something in virtue of which both George and Georg are equally morally responsible. For instance, suppose Georg sleeps in for the entire day of the killing through luck. Then there is no mental state (e.g. an intention to kill) such that both had it on the day and can be that in virtue of which both George and Georg are responsible. What grounds George's and Georg's moral responsibility?

This is similar to the Missing Option Puzzle. We've been trying to find a proposition that is true in a success world where Brenda fords the creek and a failure world where she doesn't because Black strikes her down. Similarly, the moral luck sceptic needs to find something in common between George, who successfully kills, and Georg, who doesn't because of lucky circumstances. Moreover, just as the Counterfactual Strategy appeals to a counterfactual, so does the moral luck sceptic. More precisely, the moral luck sceptic says that both George and Georg are responsible because it's true that in the right circumstances (e.g. not sleeping in) they both would have killed. This is trivially true for George and substantively so for Georg.[62]

Third, note that another consequence of the Counterfactual Strategy is in the debate between Evidential Decision Theory and Causal Decision Theory. You might object to Causal Decision Theory, with its reliance on counterfactuals (or similar machinery), because of metaphysical scruples about such apparatus, or

[62] See Zimmerman (2002:565), Enoch & Marmor (2007:420-5) and Hanna (2014:684-5) for discussion on moral luck scepticism.

because Evidential Decision Theory is more parsimonious. However, if the options are sometimes counterfactuals, then this objection is undermined to some extent – because decision theory, whether it be Evidential Decision Theory or Causal Decision Theory, needs counterfactuals to deliver sensible verdicts about what the agent ought to choose in some cases.

Now I'll put together what I've argued for in Chapters 2 and 3 with what I've argued for in this chapter. First, I will combine **Subjective Decisions** and the Counterfactual Strategy. Second, I'll use the Counterfactual Strategy to pre-empt an objection to the previous chapters.

In Chapter 2, under the assumption that the agent is not uncertain about being able to make a decision, I argued for **Subjective Actions**, which says that options are all and only the actions that the agent is certain she can make. I also proposed the sophisticated formulation of decision theory, according to which, an agent ought to do as much as she can of the best option. In Chapter 3 I refined **Subjective Actions**, interpreting "actions" as "decisions" – this is **Subjective Decisions**. Now, in this chapter, I've argued that in cases where the agent is uncertain about being able to make a decision A, the A-like option is the counterfactual *if the agent were able to decide on A, then she would decide on A*.

So I've argued for an account of options and a formulation of decision theory, both under the assumption that the agent isn't uncertain of some decision that she can make it. And I've made a proposal for what the A-like option is when an agent doubts that she can decide on A. Let me put these two thoughts together. I propose, for *all* cases, the sophisticated formulation of decision theory, and the following account of options:

> **Subjective Decisions with Counterfactuals.** An agent's options are all the decisions that the agent is certain she can make. Also, for any decision to perform some action A, if the agent is uncertain about being able to make

> that decision, the counterfactual *if the agent were able to decide on A, then she would decide on A* is also an option.

In keeping with the rest of this chapter, I understand "the agent is uncertain about being able to make a decision to A" such that the agent assigns a nonextreme credence to being able to make the decision. I call the resultant conception of decision theory, understood as including this account of options and the sophisticated formulation, *Sophisticated Subjective Decision Theory (SSDT).* This is sophisticated in two ways. First, in adopting a sophisticated formulation of decision theory. Second, in adopting a sophisticated variant of **Subjective Decisions** (namely, **Subjective Decisions with Counterfactuals**).

The Counterfactual Strategy also pre-empts an objection to what I say in previous chapters. In Brenda's case, I say that the intuition is that her doubts about being able to decide to ford the creek are relevant. The Missing Option Puzzle is the challenge of finding a ford-the-creek-like option that takes into account Brenda's doubts. I've claimed a counterfactual is a suitable option here.

However, I think that *if* there were no solution to the Missing Option Puzzle (contrary to what I argue in this chapter), then the following response would recommend itself. The intuition I noted above – that Brenda's doubts about being able to decide to ford the creek are relevant – is *mistaken*. Brenda's doubts about being able to ford the creek are *irrelevant*. This means that *decide to ford the creek* is a suitable ford-the-creek-like option in Brenda's case. I think this would have been the inevitable response if it had proved impossible to find a solution to the Missing Option Puzzle.

(I think this would have been the inevitable response because there would, as far as I can see, be only one alternative, namely, that Brenda has no ford-the-creek-like option. However, I think this is unacceptable. For if Brenda has no ford-the-creek-like option, then no matter how much she desires to get across to the other side of the creek and no matter how much she hates going home, it's false that Brenda ought to realise the ford-the-creek-like option. I think it's much more plausible (though still far from ideal) to say that Brenda's doubts about being able to decide to ford the creek are *irrelevant*, contrary to the intuition.)

Now, if there were no solution to the Missing Option Puzzle and one says that Brenda's ford-the-creek-like option is simply the decision to ford the creek, then this would affect the dialectic when we move from Brenda-type cases to cases where we assume that the agent is either certain she can or can't make a decision (that is, when we consider only the sort of cases I was considering in Chapters 2 and 3). For if Brenda's doubts about being able to decide to ford the creek *do not* rule out *decide to ford the creek* as an option, then I think we would be forced to say that an agent being *certain* that she can't make a decision does not rule out *that* decision as an option – after all, the difference between the two cases is merely the agent's *degree* of doubt. This would constitute an objection to my proposing **Subjective Decisions** in non-Brenda-type cases.

However, the Counterfactual Strategy *is* a solution to the Missing Option Puzzle, so there is no need to deny that an agent's doubts about being able to make a decision are irrelevant. In that way, the Counterfactual Strategy pre-empts an objection to the previous chapters.

Appendix 3 – Pollock (2002)

Pollock (2002) discusses a similar puzzle to the Missing Option Puzzle. Here I'll do two things. First, I'll say why the puzzle is slightly different to my puzzle. Second, and this is the main thing, I'll go through why his solution doesn't work.

Pollock takes a case like Brenda's as his starting point. Recall that Brenda doubts that she can decide to ford the creek because she thinks that Black might strike her down just as she is about to decide to ford the creek. I said that the ford-the-creek-like option can't be *decide to ford the creek* because decision theory's evaluation of it doesn't take into account Brenda's doubts about being able to decide to ford the creek. The challenge as I state it is to find a ford-the-creek-like option that is evaluated in a way that appropriately takes into account these doubts. In contrast, Pollock thinks the challenge generated by Brenda's case is to find a ford-the-creek like option such that Brenda is certain that she can realise it. In other words, Pollock upholds the Certainty Constraint even in cases when the agent is uncertain of some decision that she can make it. I don't think the puzzle (generated by Brenda's case) should be formulated like this. The immediate challenge is to find the ford-the-creek-like option that is evaluated correctly by decision theory. Requiring Brenda to be certain that she can realise the ford-the-creek-like option goes too far beyond this. The Counterfactual Strategy shows this. As we've seen, the counterfactual *if Brenda were able to decide to ford the creek, then she would decide to ford the creek* is evaluated correctly by decision theory. However, Brenda isn't certain that she can make it true. So I don't think Pollock frames the puzzle correctly – his framing rules out a perfectly good solution to the genuine puzzle.

Now let's consider Pollock's solution. Pollock proposes a solution that looks similar to the Counterfactual Strategy, but as we'll see, it is in fact closer to the disjunctive strategy discussed in section 4 – in particular, it suffers the same problem as making the EU of the ford-the-creek-like option sensitive to the utilities of the wrong worlds. In this appendix I'll outline and criticise Pollock's solution.

Pollock says that Brenda's ford-the-creek-like option is a strategic action, or what he calls a *conditional policy*. An agent adopts a conditional policy when she acts one way if some condition holds, and acts another if it doesn't. Pollock thinks that Brenda's ford-the-creek-like option is a particular type of conditional policy: where the agent acts one way if some condition holds, and *does nothing* if it doesn't. In particular, Brenda's ford-the-creek-like option is *try to ford the creek if can try to ford the creek, and do nothing if can't try to ford the creek*. "Does nothing" here is to be understood as something like "performs no action", but the precise characterisation of this won't matter.

The EU of a conditional policy *B if C, null if not-C* – where *B* labels the relevant behaviour, *C* stands for the condition, and null stands for doing nothing – can be thought of as a weighted average of the EU's of *B&C* and *null&not-C*, weighted by the probability assigned to *C* and *not-C* respectively.

Let's go through how to derive this understanding of a conditional policy's EU from what Pollock says. (This paragraph is purely exegetical and can be skipped.) Pollock sets out to define the EU of conditional policies by first defining the extent to which the agent thinks a conditional policy promotes an outcome. Let $Pr_{B\ if\ C}(O)$ measure the extent to which the agent thinks the conditional policy *B if C, null if not-C* promotes the outcome *O*. (I'm going to use Pr rather than Cr to indicate the agent's credence function, in keeping with Pollock's choice of symbols.) Pollock (p.19) stipulates the following:

$$(*)\ Pr_{B\ if\ C}(O) = Pr(C) \cdot Pr(O|B\&C) + Pr(\neg C) \cdot Pr(O|null\&\neg C)$$

Given (*), Pollock (p.19) says that the EU of a conditional policy *B if C, null if not-C* is defined in a standard way, as:

$$\sum\nolimits_{O} U(O) \cdot Pr_{B\ if\ C}(O)$$

Now to show how to derive the way I'm thinking about the EU of a conditional policy:

$$\begin{aligned}
&\sum\nolimits_{O} U(O) Pr_{B\ if\ C}(O) \\
&\quad = \sum\nolimits_{O} U(O)[Pr(C)Pr(O|B\&C) + Pr(\neg C)Pr(O|null\&\neg C)] \\
&\quad = \sum\nolimits_{O} U(O)Pr(C)Pr(O|B\&C) + \sum\nolimits_{O} U(O)\, Pr(\neg C)Pr(O|null\&\neg C) \\
&\quad = \ Pr(C)\sum\nolimits_{O} U(O)Pr(O|B\&C) + P(\neg C)\sum\nolimits_{O} U(O)\, Pr(O|null\&\neg C) \\
&\quad = Pr(C)EU(B\&C) + Pr(\neg C)EU(null\&\neg C)
\end{aligned}$$

The first equation holds by (*). The rest of the equations hold by simple rearrangements plus an appeal to the definition of EU in the last step.[63] Hence, as said above, the EU of a conditional policy *B if C, null if not-C* is a weighted average of the EU's of *B&C* and *null&not-C*, weighted by the probability assigned to *C* and

[63] I mean the definition of EU *for arbitrary propositions*, as outlined in section 2, rather than for conditional policies.

not-C respectively. (I should also say that Pollock is neutral about whether the correct framework is evidential or causal decision theory, so, for instance, the conditional probability Pr(O|B&C) is replaced by the neutral Pr(O/B&C), which might be a conditional probability or a probability of a counterfactual. In keeping with this chapter, I've interpreted the latter as a conditional probability – nothing hinges on this.)

So the EU of a conditional policy *B if C, null if not-C* is a weighted average of the EU's of *B&C* and *null&not-C*, weighted by the probability assigned to *C* and *not-C* respectively. Pollock says Brenda's ford-the-creek-like option is *try to ford the creek if can try to ford the creek, and do nothing if can't try to ford the creek*. So the EU of this is:

$$\begin{aligned}&Pr(can\ try\ to\ ford\ creek)\ EU(try\ to\ ford\ creek\ \&\ can\ try\ to\ ford\ creek\ )\\ &\qquad +Pr(can't\ try\ to\ ford\ creek)EU(null\ \&\ can't\ try\ to\ ford\ creek)\end{aligned}$$

This works well in Brenda's case. It makes the ford-the-creek-like option sensitive to the utilities of worlds (in Brenda's doxastic space) where Brenda gets struck down by Black. That's because a world where null is true, i.e. a world where Brenda performs no action, is plausibly a world where she gets struck down by Black. That means the EU of *null & can't try to ford creek* is sensitive to Brenda's doubts about being able to decide to ford the creek and hence so is the conditional the policy.

The problem, as it did for the Disjunctive Strategy, comes in Jane's case:

> **Jane's case**. One day, Jane hikes from her home and into the countryside. After a while, she comes to a raging creek which blocks her path. Jane would like to get to the other side of the creek so that she can continue her hike. She slightly prefers this to returning home and spending the rest of the day watching TV. Jane assigns 0.5 credence to being able to decide to ford the creek because she thinks Black might strike her down just as she is about to decide to ford the creek. In actual fact, Jane can both go home and ford the creek. So far this is the same as Brenda's case. However, unlike Brenda's case, Jane is also unsure that she can decide to go home. In particular, she assigns 0.8 credence to being able to decide to go home because she thinks

> that Black might strike her down just as she is about to decide to go home. As before, getting struck down by Black is a very bad outcome by the agent's lights.

It looks like there should be a ford-the-creek-like option and a go-home-like option. However, for the same reason that *decide to ford the creek* won't do for the ford-the-creek-like option, *decide to go home* won't do for the go-home-like option. So we have two missing options here. Pollock would say that the options are the following two conditional policies:

> Try to ford the creek if can try to ford the creek, and do nothing if can't try to ford the creek
>
> Try to go home if can try to go home, and do nothing if can't try to go home

And their respective EU's are:

$$Pr(can\ try\ to\ ford\ creek)\, EU(try\ to\ ford\ creek\ \&\ can\ try\ to\ ford\ creek\ ) \\ +Pr(can't\ try\ to\ ford\ creek) EU(null\ \&\ can't\ try\ to\ ford\ creek)$$

$$Pr(can\ try\ to\ go\ home)\, EU(try\ to\ go\ home\ \&\ can\ try\ to\ go\ home\ ) \\ +Pr(can't\ try\ to\ go\ home) EU(null\ \&\ can't\ try\ to\ go\ home)$$

The problem is that in some worlds (in Jane's doxastic space), where Jane gets struck down just as she is about to decide to *ford the creek*, and so where null is true, it's the case that Jane can't try to *go home*. This means that the EU of the go-home-like option is sensitive to the utilities of worlds in which Jane is about to decide to ford the creek but gets struck down – this isn't right. The same remarks apply to the ford-the-creek-like option. That is, in some worlds where Jane gets struck down just as she is about to decide to go home, and so where null is true, it's the case that Jane can't try to ford the creek. This means that the EU of the ford-the-creek-like option is sensitive to the utilities of worlds in which Jane is about to decide to go home but gets struck down – this also isn't right. So Pollock's strategy, though it works in Brenda's case, delivers the wrong options in Jane's case, and for that reason, it's not an adequate response to the puzzle.

Chapter 5 – Rational Deliberation

1. Introduction

In the previous chapters I defend Sophisticated Subjective Decision Theory (SSDT). This comprises of the following account of options:

> **Subjective Decisions**. Options are all and only the decisions that the agent is certain she can make.[64]

Additionally, SSDT construes the formulation of decision theory such that an agent ought to do *as much as she can of* the best option. So when the agent's best option is one she can realise, which is the sort of case I'll be assuming in this chapter, SSDT says that the agent ought to make a certain *decision*.

This focus on *decisions* doesn't fit well with a popular picture of rational deliberation. According to this picture, defended in Skyrms (1990), Arntzenius (2008), and Joyce (2012), the agent's credences for options change in response to decision theoretic evaluations of those options. More precisely, the options are assumed to be outer actions rather than decisions. And the rational agent increases or decreases her option-credences in light of how good or bad those options look. If an option receives a better-than-average evaluation then the agent *increases* her credence in realising that option; if an option receives a worse-than-average evaluation then the agent *decreases* her credence in realising that option. This process iterates until the agent's option-credences reach an equilibrium, a stable fixed-point, at which point the agent realises any option that is evaluated best. Here I'll assume Joyce's particular version of this picture, but nothing substantial hinges on this.

Part of the attraction of the Joycean picture is dissatisfaction with the analogue of that picture where decisions replace option-credences. On this analogue, the options are usually assumed to be outer actions but can also be decisions – the

[64] Brenda-type cases won't be relevant in this chapter so I ignore the refinement involving counterfactuals proposed at the end of the last chapter.

difference doesn't matter.[65] The agent makes a decision that is evaluated best; she subsequently becomes confident of executing that decision; she then re-evaluates her options with her updated credences, and again makes a decision that is evaluated best. This process iterates, until the moment for action comes, at which point the agent decides on whatever she happens to have decided upon. This picture is the core of theories of rational deliberation proposed by Sobel (1983) (1990) and Weirich (1998) (2004).[66] I'll discuss only the core of this picture, as elaborated in Sobel (1990), rather than any one particular fleshed-out version of it – again, nothing substantial hinges on this. Dissatisfaction with this Sobellian picture of rational deliberation emerges in Richter (1986) and Cantwell (2010).

I think *there is* good reason to be dissatisfied with the Sobellian picture, though I don't think that Richter (1986) and Cantwell (2010) have made that case. In this chapter I want to do two things. First, I want to present an objection to the Sobellian picture (section 5). This is bad news for SSDT because SSDT – with its focus on *decisions* – appears to lead to the Sobellian picture of deliberation rather than the Joycean picture. So naturally, the second thing I want to do is to respond to this objection (section 6). I will do this by proposing a more nuanced account of what epistemic rationality requires of a Sobellian deliberator. Before all of this, I will go through the Sobellian and Joycean pictures in more detail (section 2) and say why SSDT leads to the Sobellian picture (section 3).

2. More on the Pictures of Rational Deliberation

The Sobellian and Joycean pictures of rational deliberation are offered in the causal decision theory framework, where an option is evaluated by the extent to which the agent thinks the option *causes* (rather than provides *evidence* for) desirable and undesirable outcomes (see Chapter 1, section 2). The key feature of this evaluation is that there are cases where what the agent believes she will do affects the evaluation

[65] See Cantwell (2012) for this picture with decisions as options.

[66] See also Eells (1985) and Harper (1986) for sympathetic treatments.

of her options. These cases are where Sobellian and Joycean deliberation are most distinctive. One such case is as follows:

> *Death in Damascus*. Harry is deliberating about which city to flee to in the hope of avoiding Death. He knows the following about his situation. Death works from an appointment book which states time and place. An agent dies if and only if she is in the same place as Death at the time stated in the book. It is now day; Harry has an appointment at midnight. Harry has time to go to either Aleppo or Damascus by midnight; and he must go to one of these cities by midnight. The appointment book is *highly reliable*. So Harry takes his being in a city at midnight to be very good *evidence* that Death will also be there. Nevertheless, the appointment book is made up a week in advance so Harry thinks that his going to a city in no way *causes* Death to be there. What ought Harry do?[67]

If Harry goes to Aleppo, then this is evidence that Death is in Aleppo, in which case going to Damascus causes the best outcome, namely, Harry's survival. So if Harry becomes confident that he'll go to Aleppo, then on a causal evaluation, going to Damascus looks best. In contrast, if Harry goes to Damascus, then this is evidence that Death is in Damascus, in which case going to Aleppo causes the best outcome, namely, Harry's survival. So if Harry becomes confident that he'll go to Damascus, then on a causal evaluation, going to Aleppo looks best.[68]

Let's apply the Sobellian picture to this case. On that picture, the options are often assumed to be outer actions but can also be decisions – the difference doesn't matter. The important elements are that the agent makes a decision that is evaluated best; then she subsequently becomes confident of executing that decision; then she

[67] See Gibbard & Harper (1978) for the first presentation of this decision problem in the philosophical literature.

[68] Both pictures of rational deliberation assume that the agent has option-credences. That is, they assume that deliberation *does not* crowd out self-prediction. I will assume that here. See Joyce (2002) and references contained therein for discussion.

re-evaluates her options with her updated credences, and again makes a decision that is evaluated best. This process iterates, until the moment for action comes, at which point the agent decides on whatever she happens to have decided upon.

Now let's see what this entails in *Death in Damascus*. Suppose Harry starts off confident that Death is in Damascus. Then on this picture, Harry first decides on Aleppo. Then he updates his credences by becoming confident that he'll go to Aleppo. His going to Aleppo is evidence that Death is in Aleppo, so he also becomes confident that Death is in Aleppo. Now Harry re-evaluates his options and decides to go to Damascus, because that looks better in light of his updated credences. This process iterates, so that Harry's decision oscillates between a decision for Aleppo and a decision for Damascus. He will eventually enact whichever decision he happens to have made when the last available time for action comes.

Now let's look at the Joycean picture. On the Joycean picture, the options are assumed to be outer actions rather than decisions. The rational agent increases or decreases her option-credences in light of how good or bad those options look. If an option receives a better-than-average evaluation then the agent *increases* her credence in realising that option; if an option receives a worse-than-average evaluation then the agent *decreases* her credence in realising that option. This process iterates until the agent's option-credences reach an equilibrium, a stable fixed-point, at which point the agent realises an option that is evaluated best.

Applied to *Death in Damascus*, the picture is as follows. Harry evaluates his options – go to Aleppo and go to Damascus – and sees that going to Aleppo looks best so he raises his credence in going to Aleppo and drops his credence in going to Damascus. This makes him confident that Death is in Aleppo. He now re-evaluates his options and sees that going to Damascus looks best. So he raises his credence in going to Damascus and drops his credence in going to Aleppo. This continues until Harry's option-credences settle on an equilibrium – a stable fixed-point at which his evaluation of his options doesn't force a change in his option-credences. Suppose dying in Aleppo is just as bad as dying in Damascus, and that Harry's going to Aleppo is just as good evidence for Death being *there* as Harry's going to Damascus is evidence for Death being *there*. Then the equilibrium in this case is one where Harry is equally confident of going to Aleppo as he is of going to Damascus. That's because in this situation Harry would evaluate going to Damascus equally as well as going to Aleppo. Given this, Joycean deliberation does *not* require that Harry change his option-credences. Rather, this is the state of mind from which Harry should act. And he should do whatever is evaluated best, which in this case is either option.

(This is a rough outline of the Joycean picture and its application to *Death in Damascus*. In this chapter I'll be focusing on Sobellian deliberation so it will be detailed enough for what follows. However, the appendix contains a more detailed outline for those interested.)

3. SSDT and the Sobellian Picture

How exactly does SSDT lead to Sobellian deliberation? Recall that SSDT comprises of the following account of options:

> **Subjective Decisions**. Options are all and only the decisions that the agent is certain she can make.

Additionally, SSDT construes the formulation of decision theory such that an agent ought to do *as much as she can* of the best option. So when the agent's best option is one she can realise, which is the sort of case I'm assuming here, SSDT says that the agent ought to make a certain *decision*. This is crucial in what follows.

On the Sobellian picture, the agent makes a decision that is evaluated best; then she subsequently becomes confident of executing that decision; then she re-evaluates her options with her updated credences, and again makes a decision that is evaluated best. This looks like the result of iteratively applying SSDT and applying standard epistemic norms. At t1 we apply SSDT, so the agent ought to make the decision that is evaluated best. After she makes this decision, she ought to conform to standard epistemic norms, so she should update her credences by becoming confident of executing this decision. Then we *re-apply* SSDT at t2, so she ought to make the decision that is evaluated best. And so on. That is the sense in which SSDT leads to the Sobellian picture of deliberation.

In contrast, suppose we tweak SSDT by replacing **Subjective Decisions** with an account of options that says an agent's options are outer actions. An application of this decision theory would initially say that, for instance, Harry ought to go to Aleppo (assuming he starts off confident that Death is in Damascus). So it seems that Harry ought to go to Aleppo and that's the end of it – there's no implication that Harry's decisions should oscillate as they should on Sobellian deliberation. You might think that if Harry goes to Aleppo then he must decide to go to Aleppo, and at this point we can reapply this decision theory to Harry with his updated credences. However, all that would happen here is that Harry would be under two conflicting obligations: an obligation to go to Aleppo, which became binding at some earlier time, and an obligation to go to Damascus, which became binding at a later time. It's not clear what Harry ought to do here, so it's not clear that there is decision oscillation.

4. Richter and Cantwell on Sobellian Deliberation

In this chapter I'm doing two things. First, I'm objecting to the Sobellian picture (section 5). This is bad news for SSDT because SSDT – with its focus on *decisions* – appears to lead to the Sobellian picture of deliberation. The second thing I want to do is to respond to the objection (section 6). Before all of that, I want to look at three objections to the Sobellian picture *which I think don't work*. This clears the ground for a better objection in the next section. The first two objections are from Richter (1986); the third is from Cantwell (2012). Richter's objection is that Sobellian deliberation wastes time, so the objection is that it is *practically irrational.* In contrast, Richter's second objection, Cantwell's objection, and indeed my objection in the next section, all see the problem with Sobellian deliberation to be its relation to *epistemic rationality*. More specifically, each of these objections tries to show that the Sobellian deliberator is epistemically irrational. The basic idea is that there is something epistemically irrational about an agent deciding on, say, Aleppo but being prepared to evaluate her options and overturn this decision – it seems that the agent should realise that her decision will be overturned.[69]

4.1 Richter's Objection from Time-Wasting

Richter (1986) argues that Sobellian deliberation wastes time. He dramatizes this point by imagining that the agent is deliberating in a desert:

> Add to the [*Death in Damascus*] case that the man is equidistant from both cities. It will take 4 hours to travel to either city, but he must be at the gates by 6 p.m. or be shut out. It is now 10 a.m.; so the man has 4 hours before he must make a final decision. It is very hot and uncomfortable in the desert. The sooner he decides, the better, because at least he'll have that much more time for pleasure and comfort before Death takes him. The problem is that according to [the Sobellian picture], rationality requires the man to stay glued to that spot, sweating away and dithering about in a state of indecision, for the full 4 hours. The minute he tries to terminate deliberation and head off to one city, [the Sobellian picture] and rationality require that the man change his mind in favour of the other city. This position is clearly awkward. Surely our notion of rationality is such that the man can see the futility of such dithering and arbitrarily (and finally) choose one city or the other within an hour and not be irrational for doing so. (p.347)

Richter's objection to Sobel's picture is that, on that picture, Harry's decision oscillates right up until the last available time for action comes. Consequently, if Harry deliberates in the desert, then it appears that Harry irrationally dithers in an uncomfortable environment instead of simply going to a city straightaway. So Richter's objection to the Sobellian picture is that it *wastes time*, that is, it is practically irrational.[70]

[69] I should add that Richter and Cantwell see their objections as targeting CDT via the Sobellian picture. However, I'm using their objection to target SSDT via the Sobellian picture. Essentially, the combination of CDT and SSDT leads to the Sobellian picture. If one objects to the Sobellian picture, then one can use this to object to CDT or SSDT (or both). Richter and Cantwell use it to object to CDT. In contrast, I use it to object to SSDT.

[70] I note that this objection may also apply to the Joycean picture. However, it seems less pressing in that framework. For in that framework, the agent acts when she reaches an equilibrium so she doesn't necessarily deliberate *all the way* until the last moment for action comes, as she does on the Sobel picture.

I don't think this objection works. The Sobellian picture portrays the agent as making a decision that looks best, updating her credences and then making a further decision. As I understand it, this is what happens *all things being equal*. But with the introduction of the desert, all things are *not* equal. Recall that on the Sobellian picture the agent decides, updates, decides, and so on, until the last available moment for action comes. If Harry really is in a desert deliberating, then this looks like a good reason to construe the last available moment for action *early*. On this picture of rational deliberation, the decide-update-decide iterative process takes place in a given window of deliberation. When the decision needs to be made sooner rather than later, and perhaps when decisions have very high cognitive costs, then this window is very short. Richter's case is one where Harry deliberates in the desert. This is a case where the decision needs to be made sooner rather than later, so it looks like the window of deliberation is very short. If this is right, then the Sobellian deliberator does *not* waste time deliberating. In short, in precisely those circumstances where Sobellian deliberation would waste time, the window of deliberation is curtailed and the decide-update-decide iterative process takes place only in this window. So Sobellian deliberation does *not* lead to wasting time.

4.2 Richter's Objection from Epistemic Irrationality

Richter's (1986) second objection to Sobellian deliberation concerns a variant of the *Death in Damascus* case. The case is as follows:

> Two qualitatively indistinguishable clones, A and B, must decide in 20 minutes whether or not to press a large button on a console before each. They are in separate rooms and must choose independently; but both know it is virtually certain that one will choose exactly as the other chooses. If both press, each will get $10; if one presses and the other does not, each will get $100; however, if both fail to press, each loses $1000. The decision matrix, then, may be represented as follows:

Let Pi = i presses the button, then

	P_B	Not P_B
P_A	10	100
Not P_A	100	-1,000

(Richter 1986, p.348)

Richter is here assuming the options are the outer actions of pushing and not pushing the buttons rather than decisions. I'll follow him on this – nothing substantial hinges on it. Richter is also assuming that the conditional probabilities are such that Cr($P_A|P_B$)=1 and Cr(Not $P_A|$Not P_B)=1. Given this, if A assigns a credence of x>0.925 to pushing, then the causal EU of not pushing is best. And if A assigns a credence of x<0.925 to not pushing, then the causal EU of pushing is best.[71]

Now for Richter's Objection from Epistemic Irrationality. Richter considers the Sobellian picture, according to which the agent's decision would oscillate between deciding to push and deciding to not push, and says:

> This picture, however, is simply incoherent. The picture requires that A, being confident that he will push, then make a further calculation, the result of which will be that he will change his mind and incline himself to not pushing. But how can a rational person be confident he will push the button, *at the same time,* he is prepared to make, *and act on,* a further expected utility calculation that for all he knows will indicate the rationality of not pushing?! A rational person is simply not entitled to a degree of confidence of over .925 that he will push when he knows he is prepared to act on a further calculation that may well result in his not pushing. Yet a degree of confidence of over .925 is *required* in order for the decision to oscillate.
>
> So as long as A knows there is time for further deliberation, the decision cannot oscillate as depicted; he can't acquire new

[71] Why? Because if clone A assigns a credence of x to pushing then she assigns a credence of x to *if I were to push, then I'd get $10* and a credence of (1-x) to *if I were to push, then I'd get $100.* So assuming financial gain goes along with utility, the causal EU of pushing is x10 + (1-x)100. She would also assign a credence of x to *if I were to not push, then I'd get $100* and a credence of (1-x) to *if I were to not push, then I'd lose $1,000.* So the causal EU of not pushing is x100 + (1-x)(-1,000). This entails that when x>0.925, not pushing gets a better causal EU. And it entails that when x<0.925, pushing gets a better causal EU. (These numbers are correct to 3 significant figures.)

> information in a way that will overturn the previous rationality determination. But this result only leads to deeper paradox and incoherency. For A is entitled to reason through the situation exactly as we have done here, and he can know that he will enter the final calculation with a degree of confidence less than .925 that he will push. But that means that the final expected utility calculation will definitely favour the rationality of A's pushing; for at the start of the calculation he will assign B's pushing a probability of less than .925. But if this is so, A can be virtually certain, well in advance, that in the final analysis, he will push. (Richter 1986, pp.348-9)

I'm going to go through this objection carefully, so it will take a bit of time. Richter supposes that the Sobellian agent decides to push (so her credence in pushing must have been x≤0.925). Given this, there's a dilemma for the agent.

First, suppose the agent sets her credence in pushing to x>0.925. This seems strange. That's because the agent is left with a credence in pushing of x>0.925 but with a causal EU of not pushing that exceeds that of pushing. She should realise that she'll decide to not push in the next instant. In other words, it leaves her in an epistemically irrational position.

So the Sobellian agent can't rationally raise her credence in pushing to x>0.925. So let's suppose the agent sets her credence in pushing to x<0.925. The problem here is that on the resultant credences pushing will continue to look best, so (assuming that further decisions to push don't shift her credences), she won't ever revoke her decision to push and she'll end up pushing. She should realise this and be *more* confident of pushing! In other words, setting her credence in pushing to x<0.925 leaves her in an epistemically irrational position.

The upshot is that the Sobellian agent can't help but update her credences after deciding in such a way that violates an epistemic norm: after having made a decision to push, whatever she does with her credences, she does something epistemically irrational.

Let me pre-empt a worry about this objection. You might think that after deciding to push it's rational for a *myopic* agent to raise her credence in pushing to x>0.925. Richter said it was strange for the agent to raise her credence in this way because the agent should realise that she will end up deliberating in such a way that means she'll decide to not push. But this requires some foresight about how she will deliberate. So you might think that deciding to press is strange only if the agent is a *sophisticated agent* with foresight about how she will deliberate rather than a *myopic agent* with no such foresight. I think this is right. However, this does not constitute a response to Richter. Richter is formulating an objection to Sobellian deliberation, so it's enough that we find one sort of agent (in this case, a sophisticated agent) for whom Sobellian deliberation leads to epistemic irrationality. Moreover, this sort of agent is not a *hyper*-idealised agent – it's simply an agent with foresight about how she will deliberate. So I don't think the observation that Richter's objection assumes a sophisticated agent is a good response to Richter. This is important because both Cantwell's objection and my objection assume a sophisticated agent as well.

However, I do think there's a good response to Richter's objection. Richter's objection takes the form of a dilemma: it asks us to suppose that the Sobellian agent decides to push, and asks us to consider her epistemic rationality if she (i) sets her credence in pushing to x>0.925 and (ii) sets her credence in pushing to x<0.925. Clearly, there is a third option here: namely, x=0.925. Crucially, if the agent sets her credence to 0.925, then the two options receive equal causal EU. So supposing the agent sets her credence in pushing to 0.925, is there anything epistemically irrational about it?

The irrationality present when the agent sets her credence to x>0.925 is that she shouldn't be very confident of pushing if she knows she's going to reverse this decision. However, if her credence in pushing is 0.925, then she *doesn't* know that she'll reverse her decision – because both options receive equal causal EU, so she may decide on pushing or not pushing. So the irrationality in setting her credence to x>0.925 is not present in setting her credence to x=0.925. Now consider the irrationality of setting her credence to x<0.925. The problem here was that she shouldn't assign such little credence in pushing because she should realise that this will result in her retaining this decision with the result that she pushes. However, if her credence in pushing is 0.925, then she *doesn't* know that she'll stick with this decision – both options receive equal causal EU, so she may decide on pushing or not pushing. So the irrationality in setting her credence to x<0.925 is not present in setting her credence to x=0.925.

Thus there doesn't seem to be anything epistemically irrational about the agent setting her credence in pushing to x=0.925. So I think there is the following response to Richter: he hasn't shown that *whichever way* the agent updates her credences after deciding to push she is irrational, because there is nothing irrational about her deciding to push and then setting her credence in pushing to x=0.925.

4.3 Cantwell's Objection from Epistemic Irrationality

Cantwell (2010) makes a similar objection to Richter's Objection from Epistemic Irrationality, but it also appeals to an alleged intimate connection between credence and decision (which is supposed to reflect the well-discussed belief-intention bridging principles – more on this later). Cantwell outlines Sobellian deliberation in *Death in Damascus*, then he says:

> However, the situation is more complex than the story lets on. For if you expect that a decision to go to Aleppo will be revoked then you have no grounds for thinking that a decision to go to Aleppo will make it likely that you will actually go to Aleppo, and if this link is broken (so that your decision to go to Aleppo does not result in an expectation that you will go to Aleppo) then you will not regard your decision to go to Aleppo as evidence that you will go to Aleppo and so you will have no grounds for thinking that Death will have predicted that you will go to Aleppo, making the decision to go to Aleppo ratifiable (but, you believe, inefficacious). So ratifiability is regained, but at the cost of decision efficacy: the decision to go to Aleppo becomes ratifiable only because you have no reason to believe that the decision will lead to the act.
>
> Here one might raise the objection that it is not possible to form an intention to act if one does not expect that the forming of the intention will at least increase the probability that the act will be performed (compare Kavka's Toxin Puzzle above). If this is the case - the objection continues - then Death in Damascus and the Psychopath Button are not really *decision* problems: they do not involve two alternative decisions that can be made. One obvious reply is that if these scenarios do not involve decisions, they do not provide problems for a decision theory. The counter-reply is that in both Death in Damascus and the Psychopath Button the impossibility of making an

> efficacious decision is inflicted not by some physical disability or some external force but by assumptions about what constitutes practical and theoretical rationality. These assumptions can be questioned. (Cantwell 2010, p.140)

Like Richter, Cantwell says that once the Sobellian deliberator makes a decision to go to Aleppo, she *shouldn't* become confident that she'll go to Aleppo, because she should realise that she will then reverse this decision in favour of a decision for Damascus. In other words, the agent's conditional credence in going to Aleppo given that she decides on Aleppo shouldn't be high. However, whereas Richter claims there is also something epistemically irrational about the agent's conditional credence not being high, Cantwell appeals to an alleged connection between credence and decision. (Strictly-speaking, he appeals to the connection between *intention* and credence, but Cantwell is assuming that decisions on the Sobellian picture just are intentions. I will follow him in this and use "decision" and "intention" interchangeably.) In particular, he says that it is impossible to form an intention to go to Aleppo if one doesn't expect *that* intention to be effective.

The upshot of the objection can be put as follows. If the agent is confident of going to Aleppo given that she decides on Aleppo then she is epistemically irrational. The same goes for Damascus. If she is epistemically *rational*, then she doesn't face a decision problem – because if she is epistemically rational, then she is not confident that her decisions are efficacious, so she can't make a decision.

Cantwell appeals to the idea that it is epistemically irrational for the agent to assign a conditional credence of over 0.5 to going to Aleppo given that she decides on Aleppo – for then the decision would trigger a credence of over 0.5 in going to Aleppo which in turn would trigger that decision being overturned. A conditional credence of 0.5 is fine in this respect – because that would not be guaranteed to lead to the decision being overturned (both options would have equal causal EU in this case). Cantwell tries to find something problematic about the agent having a conditional credence of 0.5 or less in going to Aleppo given that she decides on Aleppo by appealing to the credence-intention link. He relies on the claim that such an agent *can't* intend to go to Aleppo. I'm not convinced that this claim is true. The agent (in such a situation) might be certain that if she intended to go to Aleppo when the time for action comes, then she would go to Aleppo. She does not have to think of her intention as a useless causally inert mental states. Rather, it is a state that will play a vital role in her going to Aleppo if she doesn't reconsider it, but, unfortunately, she thinks that she might reconsider it. In this sort of situation, it isn't obvious to me that the agent can't form the intention to go to Aleppo.

Nevertheless, bridging principles between intentions and doxastic states have been well discussed so let's turn to these to see if they help out Cantwell. The debate on bridging principles is cast in terms of belief (that is, all-out-belief) rather than credences.[72] Now, there is a fairly weak belief-intention link which is popular, namely, that if an agent intends to A then she believes she might A (or: if an agent intends to A, then she does not believe that she won't A).[73] However, Cantwell needs more than this because he needs to say that there is something problematic about the agent who intends to go to Aleppo and then assigns 0.5 or less credence to going to Aleppo. But if such an agent assigns, for instance, 0.5 credence to going to Aleppo, then she believes that she might go to Aleppo, so, at least as far as this fairly weak belief-intention link goes, there is nothing wrong with such an agent. So Cantwell needs something stronger.

There is the following belief-intention link which has supporters: if an agent intends to A, then she believes she will A.[74] If an agent assigns 0.5 or less credence to going to Aleppo, then (I would say) she doesn't believe that she will go to Aleppo, so this link would give Cantwell what he wants. But this alleged link between belief and intention is controversial. For instance, Bratman (1987, §3.4.2) argues that an

[72] The link may be metaphysical – that intention involves belief – or normative – having an intention normatively requires having the corresponding belief. Cantwell obviously has the former in mind – for he says it's *impossible* to form the intention without the corresponding belief. Nevertheless, either sort of link would be enough for Cantwell so I don't distinguish them.

[73] See Bratman (1987, §3.4.2)

[74] See Harman (1976) (1986) and Velleman (1989) for this view. This strong belief-intention link is popular amongst cognitivists about norms on intention. That's because Cognitivists think that norms on intentions derive from theoretical norms on beliefs. And the strong belief-intention link makes it easier to derive the norms on intentions from those norms on beliefs.

agent might intend to stop at a bookstore on the way home but, knowing that she is absentminded and tends to go into auto-pilot when returning home, she is agnostic about whether she will actually go to the bookstore. The agent does not seem irrational here.[75] So the literature on belief-intention bridging principles, although it shows that there is a precedent for the sort of belief-intention link that Cantwell needs, doesn't show that there is anything like unanimous support for that principle.

In short, Cantwell relies on a controversial claim, namely, that an agent can intend to go to Aleppo only if she is more than 0.5 confident that she will go to Aleppo given that she intends to go to Aleppo. Considered by itself, it's not clear that this is true. Moreover, well-discussed belief-intention bridging principles (which you might think support this claim) are either too weak or controversial. For that reason, I don't think Cantwell has provided an objection here to Sobellian deliberation.

5. A New Objection from Epistemic Irrationality

In this section I'll present a better objection to the Sobellian picture. Like Richter's second objection and Cantwell's objection, the objection considers the Sobellian picture for a sophisticated agent, that is, an agent with foresight about how she deliberates. And like both of these objections, the objection says that such a Sobellian agent is forced into a kind of epistemic irrationality. I argue for this by adding some detail to the Sobellian picture. In particular, I say that an agent re-evaluates her options *only if* her credences change. With this addition in hand, I say that whatever credence she assigns to her decision to go to Aleppo being effective, she is irrational. If she thinks it is effective, then she should realise that if she makes the decision, then she will become confident that she will go to Aleppo, and so her decision will be unstable. So it turns out that she *shouldn't* view her decision as efficacious. In contrast, if she thinks her decision to go to Aleppo is causally inert,

[75] See also Holton (2008) for scepticism about the strong belief-intention link.

then she should realise that if she makes the decision, then she will retain this decision and end up going to Aleppo. So it turns out that she *shouldn't* view her decision as causally inert. In short, the sophisticated Sobellian deliberator is forced into epistemic irrationality. In the remainder of the section I'll go through this carefully.

Suppose that Jim faces a *Death in Damascus* case:

> *Death in Damascus*. Jim is deliberating about which city to flee to in the hope of avoiding Death. He knows the following about his situation. Death works from an appointment book which states time and place. An agent dies if and only if she is in the same place as Death at the time stated in the book. It is now day; Jim has an appointment at midnight. Jim has time to go to either Aleppo or Damascus by midnight; and he must go to one of these cities by midnight. The appointment book is *highly reliable*. So Jim takes his being in a city at midnight to be very good *evidence* that Death will also be there. Nevertheless, the appointment book is made up a week in advance so Jim thinks that his going to a city in no way *causes* Death to be there. What ought Jim do?

Crucially, suppose the following about Jim:

> *50-50*. He starts off 50-50 on whether he'll end up in Aleppo or Damascus.
>
> *Sobellian Deliberation*. Jim is a Sobellian deliberator. In the case of *Death in Damascus* this involves the following. He makes a decision that maximises causal EU and then updates his credences. *If his credences change in a significant way,* then he will re-evaluate his options, possibly overturning his prior decision. In particular, he will again make a decision that maximises causal EU. This continues until the time for action comes, at which point Jim does whatever he happens to have decided on at that time.
>
> *Sophistication*. Jim knows that he's a Sobellian deliberator.

Sobellian Deliberation is a substantial fleshing out of what Sobellian deliberation entails in *Death in Damascus*. The important addition (to the basic story given earlier) is that the agent re-evaluates her options *only if* her credences change in a significant way. I like this addition because it makes the Sobellian deliberator more realistic. The agent re-evaluates her options only when things haven't turned out as she expected; so she doesn't waste cognitive resource by evaluating options, when that re-evaluation would lead merely to reinstating her previous decision.

Now consider the other two assumptions: 50-50 and Sophistication. My objection to the Sobellian picture will be that Jim is forced into epistemic irrationality. So it is important that these assumptions do not make Jim some sort of fanciful agent – Jim needs to be an agent we care about. I think the two assumptions are innocuous. *50-50* seems reasonable because Jim is, after all, deliberating about whether to go to Aleppo or Damascus, so splitting his credence evenly amongst these eventualities is common sense (all things being equal). *Sophistication* makes the agent aware that she is a Sobel deliberator. The proponent of Sobel deliberation thinks it describes a rational process of deliberation, so surely she thinks that there are agents who are aware that they deliberate by this method.

I will argue that Jim is forced into epistemic irrationality – that is, I will argue that whatever conditional credence Jim assigns to going to Aleppo given that he decides on Aleppo, he is epistemically irrational. (A parallel argument would show the same for his conditional credence in going to Damascus given that he decides on Damascus.) I will label Jim's conditional credence in going to Aleppo given that he decides to go to Aleppo Cr(A|dA). There are three possibilities: Cr(A|dA) is greater than 0.5; Cr(A|dA) is equal to 0.5; and Cr(A|dA) is less than 0.5. Let's go through each possibility in turn.

First, suppose that Cr(A|dA) is *greater than* 0.5. The problem here has already been explained by Richter and Cantwell, but let me go through it carefully. This conditional credence should be responsive to what Jim thinks will happen if he decides on going to Aleppo. The problem is that if Cr(A|dA) is greater than 0.5, then it is *not* a reflection of what Jim thinks will happen if he decides on Aleppo. For by *Sophistication*, Jim knows the following. If he decides on Aleppo, then he'll assign a credence of going to Aleppo greater than 0.5. This will force him to reconsider his decision and decide on Damascus. Given Jim knows this, Jim shouldn't be any more confident of going-to-Aleppo-given-that-he-decides-on-Aleppo than he is of going-to-Damascus-given-that-he-decides-on-Aleppo. In other words, if Cr(A|dA) is greater than 0.5, then it is irrational.

Second, suppose that Cr(A|dA) *is* 0.5. As before, this conditional credence should be responsive to what Jim thinks will happen if he decides on going to Aleppo. The problem is that if Cr(A|dA) is equal to 0.5, then it is *not* a reflection of what Jim thinks will happen if he decides on Aleppo. For, by *Sophistication*, Jim knows that if he decides on Aleppo, then he'll assign a credence of 0.5 to going to Aleppo. By *50-50*, this means Jim's credences in going to Aleppo and Damascus, and hence his credences in Death's location, will remain the same after making the decision. But then by *Sophistication*, Jim knows that he will not change his prior decision, because his credences haven't changed in a significant way. But if so, then Jim knows that he will end up going to Aleppo! So he should be very confident that he'll go to Aleppo given that he decides to go to Aleppo. That is to say, if Cr(A|dA) is equal to 0.5, then it is irrational.

Now for the final possibility: suppose Cr(A|dA) is *less than* 0.5. This is somewhat strange because this would indicate that Jim thinks his decision to go to Aleppo makes it *less likely* that he'd actually go to Aleppo (recall that Jim starts off 50-50 on whether he'll go to Aleppo or Damascus). There is question here over whether Jim can even make a decision to go to Aleppo given this conditional credence – can you genuinely decide to do something that you think would *hinder* your chances of doing it?[76] Regardless, there is the following problem. As before, the conditional credence should be responsive to what Jim thinks will happen if he decides on going to Aleppo. The problem is that if Cr(A|dA) is less than 0.5, then it is *not* a reflection of what Jim thinks will happen if he decides on Aleppo. For by *Sophistication*, Jim knows the following. If he decides on Aleppo, then he'll assign a

[76] Note that this is a decision-credence link that is much weaker than the one Cantwell needs.

credence of going to Aleppo of less than 0.5. This will force him to reconsider his decision. However, this will simply reinstate his decision to go to Aleppo because going to Aleppo looks even better now than it did before. After this, his credences won't change (because he hasn't made a new decision), so he'll stick with this decision and end up going to Aleppo. Given this situation, Jim should be much more confident of going to Aleppo. In other words, if Cr(A|dA) is less than 0.5, then it is irrational.

So whatever conditional credence Jim assigns to going to Aleppo given that he decides on Aleppo, it is epistemically irrational. Jim's self-awareness as a Sobellian deliberator forces him into epistemic irrationality.

That concludes my Objection from Epistemic Irrationality to Sobellian deliberation. It is an objection to that picture of deliberation applied to a sophisticated agent, that is, an agent who is aware that she is a Sobellian deliberator. The objection is that a self-aware Sobellian deliberator will be forced into epistemic irrationality. In the next section I'll reply to this objection.

6. A Reply to the Objection from Epistemic Irrationality

We've just looked at the Objection from Epistemic Irrationality. Roughly, the objection is that Jim, a self-aware Sobellian deliberator, is forced into epistemic irrationality: whatever credence he assigns to going to Aleppo given he decides to go to Aleppo, he is irrational. Ultimately, I think this objection fails, and this is what I'll argue in this section. I think it fails because the objection is an artefact of an overly simple view of what epistemic rationality requires of the sophisticated Sobellian deliberator. In particular, I think that when Jim decides on Aleppo, becomes confident that Death is in Aleppo, and then re-evaluates his options, the assumption that his credences remain unchanged for the purposes of this re-evaluation is mistaken. I think that his credences ought to reset to what they were *before* deliberation. That's because his re-evaluation means that his prior decision is no longer evidence for what he'll do. So he has the same evidence as he had before his decision to go to Aleppo. In what follows I'll go through this response in more detail.

The Objection from Epistemic Irrationality says that Jim, an agent who knows he's a Sobellian deliberator, can't be rationally confident that he'll go to Aleppo given that he decides to go to Aleppo. That's because he should realise that if he were to decide to go to Aleppo (with such confidence), then the following would happen:

> He would become confident that Death is in Aleppo. He would then re-evaluate his options. Crucially, it is assumed that his credences remain the same for the purposes of this re-evaluation – that is, it is assumed that his credence in Death being in Aleppo is high. Given this, the upshot of his re-evaluation would be a decision to go to Damascus.

If Jim realises that the above would happen, then (so the objection goes) Jim can't be rationally confident that he'll go to Aleppo given that he decides to go to Aleppo. The objection from epistemic irrationality *also* says that it is irrational for Jim *not* to be confident in going to Aleppo given he decides to go there. The upshot is that there is no way for Jim to be epistemically rational with respect to his conditional credence for going to Aleppo given that he decides on Aleppo.

In response, I propose that when Jim re-evaluates his options (having decided on Aleppo and become confident that Death is in Aleppo), his credences *do not* remain the same. In particular, his credences ought to reset to what they were before deliberation – that is, to what they were before his decision to go to Aleppo. That's because his re-evaluation means that he sees his prior decision as no longer effective. If he considers his previous decision as ineffective, then his evidence is now the same as what it was before deliberation, so his credences ought to reset to what they were before deliberation.

In the formal framework, the reset is easy to model. Here's how I see it working. When Jim decides on Aleppo and updates, he updates by Jeffrey conditionalization over the partition *I go to Aleppo* and *I go to Damascus*, becoming confident of *I go to*

Aleppo.[77] When Jim comes to re-evaluate his options, Jim simply initiates the reverse of this update: that is, he Jeffrey conditionalizes over the partition *I go to Aleppo* and *I go to Damascus*, setting his credences in these propositions to what they were before. The result of this process is that his credence function resets to what it was before any deliberation.[78]

The upshot is that if Jim were to decide to go to Aleppo, then although he would become confident that Death is in Aleppo and then re-evaluate his options, his re-evaluation would reset his option-credences and in turn his credences for Death's location. That means going to Damascus wouldn't look best, so Jim wouldn't necessarily reverse his decision and decide on Damascus. Hence his confidence in going to Aleppo given that he decides on Aleppo is not epistemically irrational. This disposes of the Objection from Epistemic Irrationality because the agent *isn't* forced into epistemic irrationality by whatever conditional credence she assigns to going to Aleppo given that she decides to go there. In particular, it is not epistemically irrational for her to be confident of going to Aleppo given that she decides on Aleppo.

That concludes my reply to the Objection from Epistemic Irrationality. Finally: what does Sobellian deliberation look like given this reply? Note that e*x hypothesi* Jim starts off 50-50 on whether he'll go to Aleppo or Damascus. So on Sobellian deliberation, it is permissible for him to decide on either Aleppo or Damascus. Let's say he decides on Aleppo (the same considerations will apply to Damascus). Then Jim will update his credences, becoming confident of going to Aleppo. I say that when he re-evaluates his options, his credences reset, so it will be permissible for him to

[77] What is Jeffrey conditionalization? Let Cr_p be the agent's prior credence distribution; let Cr_f be the agent's final credence distribution. Let $E_1 \dots E_n$ form a partition, and assume that each member of this partition has a positive prior credence with respect to Cr_p. Then an agent Jeffrey conditionalizes over $E_1 \dots E_n$ just when the following holds for all propositions H: $Cr_f(H) = \sum_{i=1}^{n} Cr_p(H|E_i)Cr_f(E_i)$.

[78] I must assume that Jim doesn't become certain that he would go to Aleppo when he decides on Aleppo, otherwise conditional probabilities on *I go to Damascus* are not well-defined for the alleged reset. This seems reasonable given that decisions might be reconsidered in Sobellian deliberation.

decide on Aleppo (again) or decide on Damascus. Whether he decides on Aleppo or Damascus, when he next re-evaluates his options, it will again be permissible for him to decide on Aleppo or Damascus. And so on.

Now you might worry that this elaboration of the Sobellian picture is contentious. For Sobel (1983) (1990) and Weirich (1998) (2004) make decision oscillation a key part of their theories of rationality. But on this elaboration, there is no decision oscillation. (In fact, this elaboration does allow some sort of decision oscillation. For an agent might make different permissible decisions at different times. But I take it that this isn't decision oscillation in any interesting sense.)

However, when correctly understood, my reply allows for decision oscillation. The elaboration of the Sobellian picture I gave immediately above is what happens *when the sophisticated agent* deliberates. It's because Jim is a sophisticated agent that epistemic rationality demands of him that he resets his credences. I haven't said anything about Sobellian deliberation *for the myopic agent* facing *Death in Damascus* – that is, for the agent without foresight about how she will deliberate. It is consistent with what I say here that Sobellian deliberation for the myopic agent proceeds as Sobel et al. think it does, i.e. the agent's decision oscillating until the final moment for action comes. So my reply to the Objection from Epistemic Irrationality *does not* contentiously lead to a dearth of decision oscillation for the Sobellian picture.

7. Conclusion

In the previous chapters I defend Sophisticated Subjective Decision Theory (SSDT). In the sort of case I've considered in this chapter, SSDT says that the agent ought to make a certain *decision*. This focus on *decisions* leads to the Sobellian picture of rational deliberation. According to this picture, the agent makes a decision that is evaluated best; she subsequently becomes confident of executing that decision; she then re-evaluates her options with her updated credences, and again makes a decision that is evaluated best. This process iterates, until the moment for action comes, at which point the agent decides on whatever she happens to have decided upon. This picture is the core of theories of rational deliberation proposed by Sobel (1983) (1990) and Weirich (1998) (2004).

I presented an objection to the Sobellian picture – the Objection from Epistemic Irrationality. Jim, a self-aware Sobellian deliberator, is forced into epistemic irrationality: whatever credence he assigns to going to Aleppo given he decides to go to Aleppo, he is irrational. This is bad news for SSDT because SSDT leads to the Sobellian picture. So in this chapter I've responded to that objection.

The Objection from Epistemic Irrationality relies on a certain conception of epistemic rationality, namely, that an agent's re-evaluation of her options generates no epistemically significant information. This is why it is irrational for Jim to think of his decision to go to Aleppo as efficacious – for if he makes that decision, he will become confident that Death is in Aleppo, and then when he re-evaluates his options, the decision is bound to be reversed. On a more nuanced conception of what epistemic rationality requires, the agent who makes a decision, updates, and then re-evaluates her options, also *resets* her credences to cancel out the effect of the post-decision update. Given this, there is nothing wrong with Jim thinking that his decision to go to Aleppo is effective. For if he were to make a decision to go to Aleppo, then this decision isn't bound to be reversed. For as soon as he re-evaluates his options, his credences reset to whatever they were prior to his decision to go to Aleppo, so deciding to go to Aleppo will again be permissible.

Appendix 4 – More on the Joycean Picture

I contrast Sobellian deliberation with a picture of rational deliberation defended by Skyrms (1990), Arntzenius (2008) and Joyce (2012). I adopt Joyce's version of this picture, but nothing substantial hinges on this for the purposes of the chapter. Below, I'll go through the Joycean picture in detail and show what it entails in *Death in Damascus*.

The Joycean picture is entailed by three constraints. Let's go through each of the three constraints.

First constraint: *Current Evaluation*

At any one time, an agent's evaluation of her options should go by the expected causal impact of the action as determined by her current beliefs (p.126). In other words, an agent should always evaluate her options by their causal EU.

Second constraint: *Seek the Good*

The second constraint is that the agent should change her option-probabilities at t_{n+1} to reflect her evaluation at t_n. More precisely, an agent should "seek the good".[79] The intuitive idea is simple: probabilities for options that are evaluated well at t_n should be increased at t_{n+1}; probabilities for options that are evaluated poorly at t_n should be decreased at t_{n+1}. For a formal explication of this, let the *status quo at* t_i be the weighted average of the causal EU's at t_i of all the options, weighted by the agent's probability at t_i for doing them. Then an agent seeks the good when her changes to her probability for an option A_1, $Cr(A_1)$, reflect the difference between the causal EU of A_1 and the status quo at t_n. The change should reflect the difference in the following way: $Cr(A_1)$ is increased if the causal EU of A_1 is above the status quo, $Cr(A_1)$ is decreased if the causal EU of A_1 is below the status quo, and $Cr(A_1)$ is left unchanged if the causal EU of A_1 equals the status quo. Of course, if $Cr(A_1)$ is already zero it shouldn't be decreased, and if it is already one, then it shouldn't be increased (pp.132-133).

The motivation for *Seek the Good* is that if an agent evaluates an action A_1 better than the status quo she'll find A_1 more desirable than not doing A_1 (or she'll be more "inclined" or "favourable" towards A_1). Given this, and given that she regards

[79] This constraint originally comes from Skyrms (1990, p.30).

herself as "free to do what she wants", the more confident she should be that she'll do A_1 (p.128).

It's worth pointing out a consequence of *Seek the Good.* In changing her options-credences the agent may as a result change her views about the causal structure of the world. This will happen in cases like *Death in Damascus* where what the agent will do is evidence for the causal effects of her options.

Third Constraint: *Full Information*

Notice that the first two constraints say nothing about how the agent should act. They just say how the agent should evaluate her options and change her credences in response to this. *Full Information* governs action:

> You should act on your time-*t* utility assessments only if those assessments are based on beliefs that incorporate all the evidence that is both freely available to you at *t* and relevant to the question about what your acts are likely to cause. (p.127)

In other words, agents should act on an evaluation E only when E is based on beliefs that incorporate all relevant and freely available information. You should act on E in the sense that you should perform an option with maximal causal EU according to E. The motivation for this constraint is that agents should always gather all the relevant and freely available information before acting. Joyce likens violating this constraint to "taking a hit in Blackjack without peeking at your hole card" (p.129).

The effect of this constraint is that agents can act on an evaluation E only when they are in a certain sort of equilibrium viz. when E doesn't force a change in the agent's views about the causal structure of the world. That's because information about the causal structure of the world is relevant information. So if an evaluation E forces a change in the agent's views about the causal structure of the world (and is hence information about the causal structure of the world), E is clearly not based on beliefs incorporating all the relevant and (now) freely available evidence, namely, E itself.

That concludes Joyce's three constraints on agents. One can see that in cases like *Death in Damascus* these constraints lead to a distinctive form of deliberation. That's because, in these cases, by *Current Evaluation* the agent evaluates her options; then by *Seek the Good* she changes her option-credences and in consequence her views about the causal structure of the world; then by *Full Information* she can't act on her evaluation but must instead re-evaluate her options; this process then repeats. As Joyce says, "the agent's reasoning is a kind of feedback loop in which she revises her beliefs in light of varying assessments of the causal efficacy of acts, and then revises her assessments of the causal efficacy of acts in light of her varying beliefs" (p.133).

Let's apply the Joycean picture to a run-of-the-mill decision problem. Consider a case where what the agent will do *does not* provide evidence for her options' causal efficacy. Then the agent evaluates her options at t_0 by *Current Evaluation*. She then changes her option-credences by *Seek the Good*. Because the options do not provide evidence for their causal efficacy, this change does not force a change in her views about the causal structure of the world. That means she can act on her t_0 evaluation in accord with *Full Information*. So this is one way in which the agent can be in the sort of equilibrium state required for action viz. her options do not provide evidence about their own causal efficacy.

Now let's consider where Joyce's constraints lead in *Death in Damascus*. Recall that when Harry is confident that he will go to Aleppo then going to Damascus looks best and when he is confident that he will go to Damascus then going to Aleppo look best. Suppose Harry starts off confident that he will go to Aleppo. Then by *Current Evaluation* Harry evaluates going to Damascus better than going to Aleppo. By *Seek the Good* Harry increases his confidence for going to Damascus, and, because going to Damascus is evidence that Death will be in Damascus, he increases his confidence that Death is in Damascus. By *Full Information* Harry doesn't act on his first evaluation because whether he will go to Damascus is relevant to the causal effects of his options. Instead, Harry will re-evaluate his options. As Harry repeats this process, he'll get more and more confident that he'll go to Damascus until he starts to become confident that Death is in Damascus. At this point his next evaluation, by *Current Evaluation*, deems *going to Aleppo* as best. By *Seek the Good* Harry increases the credence he assigns to going to Aleppo and hence to Death being in Aleppo. By *Full Information*, Harry doesn't act on his first evaluation but re-evaluate his actions. As Harry repeats this process, he'll get more and more confident that he'll go to Aleppo until he becomes confident that Death is in Aleppo. At this point his next evaluation, by *Current Evaluation*, deems *going to Damascus* as best and we are back to square one. You might think that Harry's deliberation oscillates like this indefinitely. In order to avoid this, Joyce proposes that agents should adopt credences so that she is in an equilibrium state from which she can act without violating *Full Information*:

> This is the unique stable point of deliberation, the only point at which all available information about the causal properties of acts has been taken into account. Given *Full Information*, this is *the* state that you should use when assessing causal expected utilities for purposes of action. (p.134)

This happens when Harry assigns a probability of 0.5 to Death being in Aleppo and 0.5 to Death being in Damascus. For then both options get the same causal EU and this *won't* subsequently force a change to the agent's credences by *Seek the good*. Because the options get the same causal EU at this point, Harry is permitted to do either action. As Joyce puts it, Harry can't go wrong whatever he does (p.134).

Closing Summary

This thesis is about decision theoretic options. I assume that an account of options generates a set (or possibly multiple sets) of options for an agent at the time she faces a choice, t_c – that is, the time before she's done anything, or the time at which she might ask herself "what ought I do?" (see Ch.1, sec.3). In that vein, I propose the following account of options:

> **Subjective Decisions with Counterfactuals.** An agent's options are all the decisions that she is certain she can make. Also, for any decision to perform some action A, if the agent is uncertain about being able to make that decision, the counterfactual *if the agent were able to decide on A, then she would decide on A* is also an option.

"Uncertainty" is understood such that an agent is uncertain about being able to make a decision when she assigns a nonextreme credence to being able to make it. I propose this account of options along with a sophisticated formulation of decision theory, according to which, decision theory says that an agent ought to do as much as she can of the best option. This package of views is Sophisticated Subjective Decision Theory (SSDT) (see Ch.4, sec.8). This is sophisticated in two ways. First, in adopting a sophisticated formulation of decision theory. Second, in adopting a sophisticated variant of **Subjective Decisions**.

I assume that every action (including a decision) is associated with a specific time (see Ch.1, sec.3). This is the time at or over which the action is performed. The relevant decisions are decisions made *immediately*. As I have previously put it, they are decisions made at t_i, which is the time immediately after the time the agent faces a choice. So an agent might have as an option *decide-at-t_i to raise my arm*. The relevant decisions are made (and completed) immediately, so they are minimal rather than extended actions (see Ch.3, sec.1).

The relevant ability in my account of options is a specific ability rather than a general ability or a potentiality (see Ch.1, sec.3). It is irrelevant whether it is a synchronic or diachronic ability, for decisions are minimal actions, and for minimal actions, these two accounts of ability come to the same (see Ch.3, sec.2). The ability is indexed to t_c, the time at which the agent faces her choice (see Ch.2, sec.5). So you have as an option a decision to perform some action A, if you are certain that you can-*at-t_c* make the decision for A. The content of a decision is another action, so there also corresponds a time to this action. There is no restriction on the time associated with this action. I might, for instance, have as an option a decision to go on holiday *next summer*.

In the normal case, where the agent is not uncertain of some decision that she can make it, options are decisions on SSDT. This has a number of consequences (see Ch.3, sec.5). In particular, contrary to our ordinary ways of talking, it is not outer actions that are rational or irrational, but *decisions* for outer actions. Moreover, an agent's predicted future irrationality is always relevant. When Professor Procrastinate thinks that if he accepts the invitation to write the chapter, he will unwisely procrastinate and fail to write the chapter, this is relevant to determining what he ought to do. My account has this consequence because it makes the options minimal actions. They are not extended actions whose decision theoretic evaluations would ignore the possibility that the agent irrationally starts but fails to complete the extended action

An important component of my account is that, in Brenda-type cases, when the agent is uncertain about whether she is able to make a decision to (say) ford a creek, then the counterfactual *if the agent were able to decide to ford the creek, then she would decide to ford the creek* is an option. This feature of my account makes it applicable to agents who are unsure of their decision-making abilities, and it resolves an independent puzzle about what the agent's options are in such cases. The appeal to counterfactuals has a number of interesting consequences. In particular, it parallels the moral luck sceptic's appeal to counterfactuals as that in virtue of which moral agents are responsible (see Ch.4, sec.8).

Another crucial feature of my account of options is that it is subjective: options are determined from the agent's perspective rather than by attention to her actual abilities. Decision theoretic evaluation is subjective: it evaluates options in light of the agent's beliefs and desires. My account entails that decision theoretic options are *also* subjective.

However, I do not conceive of decision theory as a *completely* subjective affair. For SSDT entails the sophisticated formulation of decision theory. This makes the agent's actual abilities relevant. This formulation contrasts to a naïve formulation, according to which, decision theory says that an agent ought to realise the best option. The sophisticated formulation says that an agent ought to do *as much as she can of* the best option. So if an addicted smoker's best option is a decision to stop smoking, and the smoker can't make this decision, then I say that there is *nothing* that the agent ought to do. (There will, however, be things that the agent ought not do. She ought not do any option that is sub-maximal. See Ch.2, sec.5.)

I think the relevance of the agent's actual abilities to decision theory's prescription makes a lot of sense. The principle that "ought" implies "can", which I have defended, means that theories of what an agent ought to do must make contact with reality somewhere (see Ch.2, sec.5). I propose that this is done *not* when determining the agent's options, but when determining what an agent ought to do once her options are settled. In other words, when using decision theory to generate a verdict about what an agent ought to do, it's at the *last* stage that the agent's abilities are relevant. Determining an agent's options and determining their decision theoretic evaluations are done without recourse to the agent's actual abilities. It's only when a prescription needs to be determined from these things, that the agent's actual abilities are relevant.

The argument for SSDT is spread across chapters 2-4. (Chapter 5, by contrast, picks up on a consequence of SSDT for pictures of rational deliberation.) My strategy for arguing for SSDT is to separate the relevant cases into two piles. First, cases where, for each decision, the agent is either certain she can make it or certain that she can't make it – these are the normal cases. Second, cases where the agent is uncertain about being able to make a decision – these are the Brenda-type cases. The motivation for separating these two sorts of cases is that Brenda-type cases pose a puzzle in addition to any puzzle posed by normal cases. The puzzle is that there appears to be a missing option – this was the Missing Option Puzzle. My argument is divided into two, corresponding to these two sorts of cases. Chapters 2 and 3 deal with normal cases; Chapter 4 deals with Brenda-type cases.

In Chapter 2, assuming normal cases, I argue that there is a puzzle for accounts of options – it appears that options must be sensitive to both subjective and objective features of the agent's decision problem. This was the Objective-Subjective Puzzle. I argued for **Subjective Actions** (options are actions that the agent is certain she can do) plus the sophisticated formulation of decision theory. The argument was that **Subjective Actions** trivially satisfies the relevant subjective constraint (namely, that the agent must be certain she can perform an action) and the resultant conception of decision theory manages to satisfy the relevant objective constraint (namely, that "ought" implies "can"). In contrast, an objective account of options, whilst it trivially satisfies "ought" implies "can", has a much harder time satisfying the relevant subjective constraint on options. So I end Chapter 2 endorsing **Subjective Actions** and the sophisticated formulation of decision theory.

Chapter 3, still assuming the normal case, took up the challenge of refining **Subjective Actions**. I proposed **Subjective Decisions** (options are decisions that the agent is certain she can make). The argument for conceiving of options as decisions is that only this sort of account generates plausible prescriptions across a range of cases. In particular, outer actions are too coarse-grained. They are compatible with making many different decisions. This leads to trouble when it matters *which* decisions the agent makes. For instance, an action might receive an okay-ish evaluation, despite a certain decision for it receiving a very good evaluation, in which case, decision theory might recommend a completely unrelated outer action. It also leads to trouble when the action is compatible with a *sequence* of decisions, where the agent makes later decisions in this sequence only if certain external circumstances come to pass. For then the outer action's evaluation might factor in the goodness of these external circumstances, generating strange verdicts about what the agent ought to do. I end Chapter 3 endorsing **Subjective Decisions** (which replaces **Subjective Actions**) and the sophisticated formulation of decision theory. This endorsement is qualified: I endorse these things *under the assumption that cases are normal.*

Chapter 4 turns to the second sort of case: Brenda-type cases. The puzzle here is the Missing Option Puzzle. Brenda doubts that she can even decide to ford a creek – what is her ford-the-creek-like option? All the *prima facie* candidates are unsuitable, because a decision theoretic evaluation of them ignores Brenda's doubts about being able to decide to ford the creek. I propose the Counterfactual Strategy. In cases where the agent is uncertain about being able to make a decision to A, the A-like option is the counterfactual *if the agent were able to decide on A, then she would decide on A*. A decision theoretic evaluation of this counterfactual *is* sensitive to Brenda's doubts about being able to ford the creek.

So to recap, I argue for an account of options, along with a formulation of decision theory, under the assumption that the agent isn't uncertain of some decision that she can make it. Considering cases where this assumption fails – Brenda-type cases – I make a proposal for what the A-like option is when an agent doubts that she can decide on A. Putting these two conclusions together generates SSDT:

- **Subjective Decisions with Counterfactuals.** An agent's options are all the decisions that she is certain she can make. Also, for any decision to perform some action A, if the agent is uncertain about being able to make that decision, the counterfactual *if the agent were able to decide on A, then she would decide on A* is also an option.

- The sophisticated formulation of decision theory: an agent ought to do as much as she can of the best option.

This package of views is endorsed *for all cases*, both normal cases and Brenda-type cases.

Bibliography

Adams, F. and Mele, A.R. 1992. The intention/volition debate. *Canadian Journal of Philosophy*. 22(3), pp.323-337.

Andrić, V. 2017. Objective consequentialism and the rationales of "'ought" implies "can"'. *Ratio*. 30(1), pp.72-87.

Anscombe, G.E.M. 1963. *Intention*. 2nd Edition. Oxford: Basil Blackwell.

Arntzenius, F. 2008. No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis*. 68(2), pp.277-297.

Austin, J. L. 1958. Ifs and Cans. *Journal of Symbolic Logic*. 23(1), pp.74-75.

Bergström, L. 1968. A discussion note on utilitarianism. *Theoria*. 34(2), pp.163-170.

Bergström, L. 1971. Utilitarianism and alternative actions. *Noûs*. 5(3), pp.237-252.

Bermúdez, J.L. 2009. *Decision theory and rationality*. Oxford University Press.

Bradley, R. 2017. *Decision theory with a human face*. Cambridge: Cambridge University Press.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, Mass.: Harvard University Press.

Bratman, M. 1992. Practical reasoning and acceptance in a context. *Mind*. 101(401), pp.1-15.

Bratman, M. 1998. Toxin, Temptation, and the Stability of Intention. In: Coleman, J.L. and Morris, C. W. eds. *Rational Commitment and Social Justice: Essays for Gregory S. Kavka*. New York: Cambridge University Press, pp.59-83.

Cantwell, J. 2010. On an alleged counter-example to causal decision theory. *Synthese*. 173(2), pp.127-152.

Carlson, E. 1999. Consequentialism, alternatives, and actualism. *Philosophical Studies*. 96(3), pp.253-268.

Copp, D. 1997. Defending the principle of alternate possibilities: Blameworthiness and moral responsibility. *Nous*. 31(4), pp.441-456.

Copp, D. 2008. 'Ought' implies 'can' and the derivation of the Principle of Alternate Possibilities. *Analysis*. 68(1), pp.67-75

Davidson, D. McKinsey, J. C. C. and Suppes, P. 1955. Outlines of a formal theory of value, I. *Philosophy of science*. 22(2), pp.140-160.

Eells, E. 1985. Weirich on decision instability. *Australian Journal of Philosophy*. 63(4), pp.473-478.

Elga, A. 2010. Subjective probabilities should be sharp. *Philosophers' Imprint*. 10.

Enoch, D. and Marmor, A. 2007. The case against moral luck. *Law and Philosophy*. 26(4), pp.405-436.

Frankfurt, H. G. 1969. Alternate possibilities and moral responsibility. *The journal of philosophy*. 66(23), pp.829-839.

Gibbard, A. and Harper, W.L. 1978. Counterfactuals and two kinds of expected utility. In: Hooker, C., Leach, J. and McClennen, E. eds. *Foundations and applications of decision theory, Volume 1*. Dordrecht, Holland: D. Reidel, pp.125-162.

Glick, E. 2012. Abilities and Know-How Attributions. In Brown, J. & Gerken, M. eds. *Knowledge Ascriptions.* Oxford: Oxford University Press, pp.120-139.

Goldman, H.S. 1978. Doing the best one can. In: Goldman & Kim, J. eds. *Values and Morals*. Netherlands: Springer, pp.185-214.

Graham, P. A. 2011a. Fischer on blameworthiness and "ought" implies "can". *Social Theory and Practice*. 37(1), pp.63-80.

Graham, P.A. 2011b. 'Ought' and Ability. *Philosophical Review*. 120(3), pp.337-382.

Haji, I. 1997. Frankfurt-pairs and varieties of blameworthiness: Epistemic morals. *Erkenntnis*. 47(3), pp.351-377.

Haji, I. 2002. *Deontic morality and control*. Cambridge: Cambridge University Press.

Hanna, N. 2014. Moral luck defended. *Noûs*. 48(4), pp.683-698.

Hare, C. 2010. Take the sugar. *Analysis*. 70(2), pp.237-247.

Harman, G. 1976. Practical Reasoning. *The Review of Metaphysics*. 29(3), pp.431-463.

Harman, G. 1986. *Change in view: Principles of reasoning*. Cambridge, MA: The MIT Press.

Harper, W. 1986. Mixed strategies and ratifiability in causal decision theory. *Erkenntnis*. 24(1), pp.25-36.

Hedden, B. 2012. Options and the subjective ought. *Philosophical Studies*. 158(2), pp.343-360.

Hedden, B. 2015. Options and diachronic tragedy. *Philosophy and Phenomenological Research*. 90(2), pp.423-451.

Honoré, A.M. 1964. Can and Can't. *Mind*. 73(292), pp.463-479.

Howard-Snyder, F. 2006. "Cannot" implies "not ought". *Philosophical Studies*. 130(2), pp.233-246.

Jackson, F. and Pargetter, R. 1986. Oughts, options, and actualism. *The Philosophical Review*. 95(2), pp.233-255.

Jackson, F. 2014. Procrastinate revisited. *Pacific Philosophical Quarterly*. 95(4), pp.634-647.

Jeffrey, R. C. 1968. Probable knowledge. *Studies in Logic and the Foundations of Mathematics*. 51, pp.166-190.

Jeffrey, R.C. 1983. *The logic of decision*. 2nd edition. Chicago: University of Chicago Press.

Joyce, J.M. 1999. *The foundations of causal decision theory*. New York: Cambridge University Press.

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

Joyce, J.M. 2012. Regret and instability in causal decision theory. *Synthese*. 187(1), pp.123-145.

Lewis, D. 1981. Causal decision theory. *Australasian Journal of Philosophy*. 59(1), pp.5-30.

Maier, J. 2015. The agentive modalities. *Philosophy and Phenomenological Research*. 90(1), pp.113-134.

Meacham, C.J. 2010. Binding and its consequences. *Philosophical studies*. 149(1), pp.49-71.

Mele, A.R. 2003. Agents' abilities. *Noûs*. 37(3), pp.447-470.

Nair, S. 2015. Moral dilemmas. [Online]. In: Routledge Encyclopedia of Philosophy. [Accessed on 31/07/18]. URL: www.rep.routledge.com/articles/moral-dilemmas.

Norcross, A. 1995. Should utilitarianism accommodate moral dilemmas?. *Philosophical studies.* 79(1), pp.59-83.

Pollock, J.L. 1983. How do you maximize expectation value?. *Nous.* 17(3), pp.409-421.

Pollock, J. L. 2002. Rational choice and action omnipotence. *The Philosophical Review.* 111(1), pp.1-23.

Portmore, D. (2017). *Opting for the Best: Oughts and Options.* [Online]. [Accessed 31/07/18]. URL: bit.ly/28XuhfT.

Prawitz, D. 1968. A discussion note on utilitarianism. *Theoria.* 34(1), pp.76-84.

Prawitz, D. 1970. The alternatives to an action. *Theoria.* 36(2), pp.116-126.

Richter, R. 1984. Rationality revisited. *Australasian Journal of Philosophy.* 62(4), pp.392-403.

Richter, R. 1986. Further comments on decision instability. *Australasian Journal of Philosophy.* 64(4), pp.345-49.

Sartre, J. 1980. *Existentialism and humanism.* London: Methuen.

Savage, J.L. 1972. The Foundations of Statistics. 2nd edition. New York: Dover Publications.

Schnall, I.M. 2001. The principle of alternate possibilities and 'ought' implies 'can'. *Analysis.* 61(272), pp.335-340.

Schwarz, W. 2017. Options and Actions. [Online]. [Accessed 31/07/18]. URL: www.umsu.de/papers/.

Skyrms, B. 1990. *The Dynamics of Rational Deliberation.* Cambridge, MA: Harvard University Press.

Slote, M. 1985. Utilitarianism, moral dilemmas, and moral cost. *American Philosophical Quarterly.* 22(2), pp.161-168.

Sobel, J. H. 1983. Expected utilities and rational actions and choices. *Theoria.* 49(3), pp.159-183.

Sobel, J. H. 1990. Maximization, stability of decision, and actions in accordance with reason. *Philosophy of Science.* 57(1), pp.60-77.

Streumer, B. 2007. Reasons and impossibility. *Philosophical Studies.* 136(3), pp.351-384.

Timmerman, T. and Cohen, Y. 2016. Moral Obligations: Actualist, Possibilist, or Hybridist?. *Australasian Journal of Philosophy*. 94(4), pp.672-686.

Vallentyne, P. 1989. Two types of moral dilemmas. *Erkenntnis*. 30(3), pp.301-318.

Velleman, D. 1989. *Practical Reflection*. Princeton: Princeton University Press.

Vranas, P. B. 2007. I ought, therefore I can. *Philosophical studies*. 136(2), pp.167-216.

Wedgwood, R. 2013. Rational 'ought' implies 'can'. *Philosophical Issues*. 23(1), pp.70-92.

Weirich, P. 1983. A decision maker's options. *Philosophical Studies*. 44(2), pp.175-186.

Weirich, P. 1998. *Equilibrium and Rationality*. New York: Cambridge University Press.

Weirich, P. 2001. *Decision space: Multidimensional utility analysis*. Cambridge: Cambridge University Press.

Weirich, P. 2004. *Realistic Decision Theory: Rules for nonideal agents in nonideal circumstances*. New York: Oxford University Press.

Whittle, A. 2010. Dispositional abilities. *Philosophers' Imprint*. 10(12), pp.1-23.

Widerker, D. 1991. Frankfurt on 'ought implies can' and alternative possibilities. *Analysis*. 51(4), pp.222-224.

Williamson, T. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

Yaffe, G. 1999. 'Ought' Implies 'Can' and the Principle of Alternate Possibilities. *Analysis*. 59(3), pp.218-22.

Yaffe, G. 2005. More on "ought" implies "can" and the principle of alternate possibilities. *Midwest Studies in Philosophy*. 29(1), pp.307-312.

Zimmerman, M. J. 1996. *The concept of moral obligation*. New York: Cambridge University Press.

Zimmerman, M. J. 2002. Taking luck seriously. *The Journal of Philosophy*. 99(11), pp.553-576.

Zimmerman, M. J. 2008. *Living with uncertainty*. New York: Cambridge University Press.