Dominance in Decision Theory

Timothy Luke Williamson

October 2021

A thesis submitted for the degree of Doctor of Philosophy at The Australian National University.

This thesis is solely the work of the author, with the exception of Chapter 1 (Sections 1.1-1.5) and Chapter 4—those chapters are drawn from papers of which I was equal first author. No part of this thesis has previously been submitted for any degree, or is currently being submitted for another degree. To the best of my knowledge, any help received in preparing this thesis, and all sources, have been duly acknowledged.

Signed: 13/10/2021

Timothy Luke Williamson

School of Philosophy, RSSS, Australian National University

Acknowledgments

I am indebted to numerous people for their personal support and professional guidance over the last few years. Indeed, I owe so many debts of gratitude that I want to begin by thanking some groups, rather than individuals. My sincere thanks go to all the visitors during various 'ANU Visitor Seasons'—thanks for the teatime chats, hallway conversations, seminars, reading groups, and above all, the exemplary model of collegiality you have provided. I also want to express my gratitude to the RSSS Admin Team—I do not always understand the system, so thanks for your patience in navigating it. And I would like to thank various anonymous referees and journal editors for their feedback on several papers—the clarity and depth of your (unpaid!) insights has improved what my work immeasurably.

Special thanks to those who have commented on specific papers or have helped me in conversation: Justin D'Ambrosio, Nicholas DiBella, Szymon Bogacz, Eleanor Boxall, Geoff Brennan, John Broome, Nicholas Carroll, Devon Cass, Peter Clutton, Garrett Cullity, John Cusbert, Domi Dessaix, Phil Dowe, Adam Elga, Edward Elliott, Melissa Fusco, Mario Guenther, Riley Harris, Brian Hedden, Jasper Hedges, Josef Holden, Bruno Ippedico, Frank Jackson, Renée Jorgensen, David Kinney, Boris Kment, James Lim, Thomas Lingafelter, Matt Kopec, Ten-Herng Lai, Hanti Lin, Kirsten Mann, Daniel Muñoz, Michael Nielsen, Daniel Nolan, Ross Pain, Susan Janet Pennings, Philip Pettit, Ignacio Quintana, Wlodek Rabinowicz, Oliver Rawle, Pamela Robinson, Wolfgang Schwarz, Toby Solomon, Nic Southwood, Katie Steele, Kim Sterelny, Daniel Stoljar, Jeremy Strasser, Yehezkel Symonds, Koji Tanaka, Teruji Thomas, Joshua Thong, Lachlan Umbers, Shang Long Yeo, Brandon Yip, Michael Yuen, Hayden Wilkinson, Patrick Williamson, and James Willoughby. I am sure that I have missed some names—if you think your name should be on the list, then it should. My time at ANU was supported by an Australian Government Research Training Program Scholarship Stipend.

I want to especially single out those friends whom I have benefited from co-authoring with: Christopher Bottomley, Christoph Lernpaß, and Alexander Sandgren. It has been a blast—you have convinced me that philosophy is best done as part of a team.

Thanks to Christopher Bottomley for permission to use our paper 'Reasonable Risk-Aversion', which forms the basis of Chapter 1, and to Alexander Sandgren for permission to use our papers 'Law-Abiding Causal Decision Theory' and 'Determinism, Counterfactuals, and Decision', which form Chapter 4. In each case, authorship was shared fifty-fifty. The writing process was highly collaborative, so it is impossible to single out specific paragraphs that are attributable to one author over another. I follow ANU's 2020 CASS HDR Guide in treating both as equivalent amount to single authored papers of half of the word count.

Chapter 2 contains material from 'Causal Decision Theory is Safe from Psychopaths', published in *Erkenntnis* (2021) by Springer. Chapter 4 includes material from 'Determinism, Counterfactuals, and Decision' published in the *Australasian Journal of Philosophy* (2021) by Taylor and Francis, and 'Law-Abiding Causal Decision Theory' published in the *British Journal for the Philosophy of Science* (Forthcoming) by University of Chicago Press.

Alan Hájek has given me more time than any student could reasonably expect. Thank you for your guidance, both in terms of how to do philosophy and, more importantly, how to approach philosophy. A true counterfactual: but for your help, I would not have completed this thesis.

The list of people who have personally supported me is longer still. I want to thank my parents, Dugald and Rose Williamson. Special thanks go to Aden Cotterill, Steve Prior, and Patrick Williamson—I have learned more than I can say from each of you. My greatest thanks go to Claire Williamson, to whom this thesis is dedicated.

This research was conducted with

Abstract

Philosophers often appeal to *dominance principles* as they try to work out what the correct decision theory looks like. In this thesis, I aim to get clear on just what dominance principles are, and the work they can be put to. I consider dominance principles for (i) agents with *risk-averse* preferences, (ii) agents facing *Newcomb's Paradox*, and (iii) agents with *cyclic* preferences.

In the case of risk-averse preferences, I motivate and defend two constraints that have been overlooked by philosophers, *Decomposability* and *Betweenness*. In the case of Newcomb's Paradox, I outline both recent challenges to and defences of Causal State-wise Dominance. I conclude that the debate over Newcomb's Paradox is still at a standoff, despite those recent interventions. In the case of cyclic preferences, I defend the rationality of cyclic preferences against various forms of Money Pump. I then consider cyclic preferences in the context of risk, in particular the tension between cyclicity and State-wise Dominance. Building on a plausible set of intuitions about cases, I develop a view that rejects State-wise Dominance, embraces Money Pumps, and sometimes recommends choosing an option that is strictly dispreferred to every available alternative. While this is a far more minimal view of dominance than many would accept, I argue that it is coherent.

My overall conclusion is that we should think about dominance principles as *domain specific* tools—tools that under the right conditions simplify reasoning and demonstrate when agents are failing to live up to their own standards. They need not be exceptionless rules to play these roles.

Introduction

0. Dominance

Many of our decisions are *hard*. We must weigh up chances of gains against chances of losses; we must judge the magnitudes of those gains and losses; then there are our risk-attitudes, the question of how (and whether) to assign precise probabilities to events, and a whole raft of complicating factors to consider. This thesis is not about those hard choices, at least not directly. This thesis is about *easy* decisions—those situations in which one choice appears to be *obviously* or *unambiguously* better than the rest. More precisely, this thesis is about dominance principles, a core set of decision-theoretic principles that capture the sense(s) in which one act can be better than another in every respect. Dominance principles are extraordinarily useful in situations where they apply—they are rational shortcuts that let us bypass complicated weighing of probabilities and values. And in situations where dominance principles do not apply directly, they are useful building blocks for a decision theory that does apply—we can assess theories based on whether they respect appropriate dominance principles in simple cases. So, dominance principles promise to make our lives easier both as decision-makers and as philosophers wanting to formulate rules to guide decision-makers.

What then is a dominance principle? That is the motivating question behind this thesis. The basic idea is straightforward enough: one choice dominates another when there is *no sense* in which the latter does as well as the former. But like many slogans, this turns out to be challenging to spell out in precise terms. As will become evident, there are many ways of precisifying the basic idea behind dominance, and we can challenge all of them. Indeed, various plausible dominance principles may conflict with each other. So, while dominance principles pick out a subclass of decisions that count as easy, saying just what those principles are is hard.

Let me be upfront about the limitations of this thesis. I narrow the focus in two ways. Firstly, I focus on dominance as it arises in the context of *instrumental rationality*. A theory of instrumental rationality tells you which means are appropriate to take to your ends, but it does not specify what those ends are. I am therefore concerned with what it means to say that one option is in every sense as good a means to your ends as another. We could, of course, ask questions about ends themselves (e.g., when there are multiple dimensions of value, in what sense can some ends dominate others?). But those would be questions for a different thesis.[1] Secondly, I focus on dominance principles as they arise in the context of *individual rationality*. A theory of individual rationality tells us what is rational for a single agent (I typically refer to them as 'you'), not what is rational for groups of agents. Again, we could ask questions about dominance in the context of social choice theory, game theory, population ethics, and so on. But those too would be questions for a different thesis.

With those restrictions in place, let me state my overall conclusion: most dominance principles are false. That is, apart from a small number of often-overlooked minimal constraints, most dominance principles admit of exceptions. Nonetheless, dominance principles are useful when they apply. They are domain-specific tools that play important theoretical and practical roles, and they need not be decision-theoretic bedrock, as many have assumed, to play those roles. Let me outline how I get to this conclusion.

0.1 Thesis Overview

In the second half of this Introduction, I survey dominance in decision theory. I distinguish three families of dominance principle, those governing *preferences*, *choices*, and *plans*, respectively. For each family, I outline some widely accepted principles and cases in which those principles have been challenged.

In Chapter 1, I turn to a dominance principle that has received little philosophical attention, *Decomposability* and the closely related *Betweenness* axiom. Decomposability- and Betweenness-satisfying theories are more permissive than orthodox Expected Utility Theory, for example in permitting the Allais Preferences. But recent departures from Expected Utility such as Lara Buchak's (2013) *Risk-Weighted Expected Utility Theory* (REU) violate Betweenness and Decomposability. I argue that we should reject REU and adopt a Betweenness-satisfying model instead. I sketch Chew's (1983, 1989) *Weighted Linear Utility Theory* (WLU), provide a novel normative interpretation of that theory, and show that the resulting view has independent advantages over REU.[2]

[1] Note also that I am happy to talk about means and ends. Contrast this with a strongly *coherentist* view of decision theory, on which rationality is simply a matter of adopting means that 'fit' or 'cohere' with each other (see the helpful discussion in Buchak 2013, Section 5.1). A guiding assumption throughout is that there are 'ends' that you care about, and your attitudes towards those ends plays a role in constraining appropriate means.

[2] Sections 1.1-1.5 of Chapter 1 are drawn from a co-authored paper with Christopher Bottomley (see Bottomley & Williamson, Manuscript). Co-authorship was equal, and I thank Chris for his permission to make use of that paper here.

In Chapters 2-4, I turn to the ongoing debate between *Causal* and *Evidential* decision theorists. The core of Causal Decision Theory (CDT) is a commitment to *Causal State-wise Dominance*, which says that one option is better than another if it does better regardless of how the world turns out to be, where 'how the world turns out to be' is defined in terms of is *causally independent* of your choice. Evidential Decision Theorists (EDT) dispense with metaphysical concepts like causation and adopt instead a purely Bayesian framework—each act is evaluated based on probabilities of outcomes conditional on your performance of that act. As a result, EDT violates Causal State-wise Dominance in the famous Newcomb Paradox.

I argue in these chapters for *biased pluralism*. Like other pluralists (e.g., Bales 2018a), I do not think that rationality itself tells us whether we should uphold or reject Causal State-wise Dominance. Nonetheless, when we ask what we want a decision theory *for*, EDT has pragmatic advantages over CDT. While these advantages do not show the causalist to be irrational, they persuade me to adopt the evidentialist standard.

In slightly more detail:

In Chapter 2, I review various standard arguments for both CDT and EDT and conclude that the debate is indeed 'hopelessly deadlocked' (Lewis 1981a, p. 5). Prominent arguments for Causal State-wise Dominance, those from *Full Information* and *Actual Value*, fail. Similarly, prominent arguments against Causal State-wise Dominance, the *Why Ain-cha Rich?* argument and various arguments from *Decision Instability*, fail. These arguments all fail for a single reason: the causalist and evidentialist adopt different notions of *independence*, or which parts of the world to hold fixed when assessing different acts. And each of the arguments I consider presupposes a view of independence that the target view already reasonably rejects.[3]

In Chapter 3, I discuss a novel class of alleged counterexamples to CDT, Arif Ahmed's (2013, 2014a, 2014b) *deterministic cases*. These cases promise to break the deadlock by showing that Causal State-wise Dominance requires you to perform acts that are patently irrational. Ahmed's first case, *Betting on Laws*, purportedly shows that CDT advises you to bet against your own credences, say by betting against the truth of some law-like statement that you are overwhelmingly confident in. Ahmed's second case, *Betting on the Past*, shows that in cases where states carry information about what you are determined to do, CDT requires you to accept bets you cannot possibly win. In response to Betting on Laws, I argue that the problem lies not with CDT but the standard approach to counterfactuals from Lewis (1973), which many couple with

[3] Chapter 2 is an expanded and updated version of Williamson (2021).

CDT. I develop an impossible worlds approach that resolves Ahmed's challenge. In response to Betting on the Past, I agree with Ahmed that CDT yields absurd verdicts in that case. But I do not think that we must reject a broadly causalist position. I propose *Selective* Causal Decision Theory (SDT), which agrees with CDT in a wide range of cases while departing from it in cases where states carry information about what you are determined to do. SDT preserves enough of the motivation behind CDT to be called causal, and it respects an appropriately modified version of Causal State-wise Dominance. So, while deterministic cases do not resolve the stalemate between causalists and evidentialists, they do force the causalist to clarify the details of their view.[4]

In Chapter 4, I argue that not all stalemates are equal. The debate between CDT'ers and EDT'ers is at a stalemate only in the sense that you cannot convince a reasonable proponent of one view to adopt the other. But, taking a cue from Horgan (1981, 2017), I note that our goal need not be to convince the proponent of one view to so change their mind—each of us might construct a decision theory that seems right to us. When constructing a decision theory, we might then ask what we want a decision theory *for* and then ask whether CDT or EDT does a better job of achieving that goal. I want a decision theory to provide *action-guiding* advice, and EDT does a better job as an action-guiding theory on two counts. Firstly, CDT advises you to do things that you will not do in virtue of following CDT. So, it is unclear how we are supposed to implement CDT's advice. Secondly, CDT's advice is sensitive to how you frame your options. So, when there are multiple, legitimate ways of framing your options, CDT delivers inconsistent verdicts. And again, it is unclear how we are supposed to act on the basis of such advice. While neither of these features need persuade the causalist to abandon their view, somebody not already committed to causalism might take them as reasons to adopt the evidentialist standard instead.

The upshot of Chapters 2-4 is that though Causal State-wise Dominance is a natural first-pass dominance principle, it is not sacrosanct. It conflicts with reasonable intuitions in Newcomb's Paradox, needs to be qualified in deterministic cases, and meta-considerations persuade me to adopt a different standard.

Chapter 5 discusses dominance in the context of cyclic preferences, for example when you prefer Abba to Bach, Bach to Chopin, and Chopin to Abba. I first consider Money Pumps, which can be construed as dominance arguments against cyclic preferences. I argue that both diachronic

[4] This chapter is a streamlined version of two co-authored papers with Alexander Sandgren (see Sandgren & Williamson 2021 and Williamson & Sandgren, Forthcoming). Co-authorship in both cases was equal, and I thank Alex for permission to use both papers here.

Money Pumps (due to Davidson et al. 1955) and synchronic Money Pumps (due to Gustafsson 2013) fail to undermine the case for rational cyclic preferences.

I then outline a paradox for cyclic preferences under conditions of risk. I present a simple case, the *Easy Lottery*, in which cyclic preferences require you to either (i) violate State-wise Dominance, or (ii) pay to relabel the tickets of a fair lottery. This case provides us with either a novel *counterexample* to cyclic preferences or a *challenge* to clarify the structure of decision theory without transitivity. I do not take a stand on whether the case is a counterexample or challenge, but I show how each possible response can be rationalised with some formal rule for comparing acts, which in turn sheds light on the normative underpinnings of non-transitive decision theories.

Chapter 6 ends on a positive note. Some think that cyclic preferences cannot guide choice, since in some situations *anything* you might choose is dispreferred to an available alternative. Building on the discussion in Chapter 5, I propose to compare acts as means to ends with Fishburn's (1982, 1984a,b) *Skew-Symmetric Bilinear Utility Theory.* I then develop a simple decision rule: choose options that are *Least Bad* (i.e., those that minimise negative utility-difference when compared to available alternatives). While this might not be the only sensible rule that allows for cyclicity, I argue that it satisfies (or is compatible with) plausible constraints on choice. Of note is that on my view *fully dominated* options—those that are strictly dispreferred to every available alternative—can be permissible. I think that this is a feature, not a bug (or if a bug, an easily remedied one).

In the conclusion, I note that the picture of dominance I end up with is far more minimal than the conventional picture. No matter. Well-known dominance principles hold in the cases you would expect—for example, when outcomes are well ordered or when agents can bind themselves to courses of action. This casts such principles in a new light: useful *domain-specific* tools that help us get clarity on the structure of instrumental rationality and that simplify choices under the right conditions. So, even if most dominance principles are false as general principles, they may still play important practical and theoretical roles.

0.2 Overviewing Dominance

Dominance-talk gets used in many ways, and it is not always clear what is going on when someone appeals to a 'dominance' principle. I here provide a bird's eye view of dominance principles and the work they have been put to. This survey is necessarily broad—every issue I discuss deserves a fuller treatment. But I hope that even a broad survey gives the reader an idea of (i) how important it is to get dominance right, and (ii) how careful we must be in endorsing some principle simply because it bears the name 'dominance'. This will also serve to locate subsequent chapters in their wider decision-theoretic context.

Before getting into the principles themselves, we need a formal framework. I adopt Savage's (1972) framework that assumes a set of *states* S, a set of *outcomes* O, and a set of acts defined as the set of possible functions $f: S \rightarrow O$. Think of each state as describing the parts of the world outside of your control—states are *independent* of your acts—and, moreover, describing the world in enough detail that each act-state pair determines a single outcome. Clearly not every Savage-act is something that you can do at will—the acts that you are in a position to perform I refer to as *options*. The outcomes describe everything you care about, and they correspond to your ends, or propositions of intrinsic concern (this means that you are indifferent between the various ways an outcome might be realised).[5] I assume that you have preferences over outcomes: $x \geq y$ says that you *weakly prefer* x to y; $x = y$ says that you are *indifferent* between x and y and holds whenever whenever both $x \geq y$ and $y \geq x$; and $x > y$ says that you *strictly prefer* x to y and holds whenever both $x \geq y$ and $\neg(y \geq x)$. I take it as given that preferences satisfy:

- **Reflexivity:** For all $x \in O$, $x \geq x$.

I typically assume that preferences satisfy both:

- **Completeness:** For all $x, y \in O$, $x \geq y$ or $y \geq x$, and
- **Transitivity:** For all $x, y, z \in O$: (i) if $x > y$ and $y > z$, then $x > z$, (ii) if $x \geq y, y \geq z$ then $x \geq z$, (iii) if $x = y$, $y = z$, then $x = z$.

I do not take Completeness and Transitivity to be necessary features of preferences. At various stages I consider dropping both, but I make it clear when I do so, so both can be assumed unless otherwise stated.[6]

[5] An alternative *state-dependent* approach would allow that your preferences over outcomes may depend on which states those outcomes are realised in (see Karni, Schmeidler, and Vind 1983 and Karni and Schmeidler 2016 for this approach). Since I take each act-state pair to determine everything you care about, outcomes correspond to what Sobel (1986) calls a *world*, though we must be careful to distinguish this from a world in the metaphysician's sense.

[6] I am broadly neutral on what precisely preferences are. Often, when one outcome is strictly preferable to another, I say that the strictly preferable outcome is better for you or what you have most reason to choose. This might be because we analyse betterness and reasons in terms of preference. Or it might be because preferences are appropriately related to what is better for you and what you have most reason to choose. For example, perhaps what you should prefer what is better for you, or perhaps preferences are a certain kind of judgement of betterness (cf. Hausman 2011, Bradley 2017, pp. 46-47), or perhaps preferences supervene on what you have most subjective

Since I am concerned with which means are appropriate to take to your ends, we need to introduce preferences over *acts*: again, $f \succcurlyeq g$ says that f is *weakly preferred* to g; indifference $f \sim g$ holds when both $f \succcurlyeq g$ and $g \succcurlyeq f$; and strict preference $f \succ g$ holds when both $f \succcurlyeq g$ and $\neg(g \succcurlyeq f)$. I refer to an act f_x that yields a single outcome x in each state as a *degenerate* act (sometimes called a *constant* act), and I assume $f_x \succcurlyeq f_y$ if and only if $x \geq y$. It will also be helpful to talk about compound acts: for $E \subset S$, define $f_E g$ to be an act that agrees with f for all $s \in E$ and with g for all $s \notin E$. Throughout I work only with finite acts—those that map each state to one of a finite number of outcomes.

In addition to acts, states, and outcomes, I assume that your partial beliefs can be represented by a subjective probability function over S, your *credence function* C. This property is sometimes called *probabilistic sophistication*. Savage's own approach was to derive your subjective probability function from preferences over acts,[7] but see Hájek (2008) and Meacham and Weisberg (2011) for prominent critiques of this approach to defining subjective probability. For my purposes, I assume the existence of a probability function and largely set aside how it is characterised and elicited.[8] I will further assume that C is defined not just on states but on acts (though such act-credences are a controversial addition to the standard Savage-framework, I accept the defences by Rabinowicz 2002 and Hájek 2016).

So, there are things you control (acts), things you cannot control (states), and things that you care about (outcomes). A decision theory takes as input your credences and preferences over outcomes and says, in light of those credences and preferences, which options you may choose.

Often, it does not matter which outcomes occur in which states, so we can talk about an act by describing the probability distribution that it induces over outcomes. In that case, we can associate f with $\alpha_1 o_1 + \cdots \alpha_n o_n$, where α_i is the total probability of the states such that $f(s) = o_i$. Some argue that we can identify each act with the probability distribution it induces over

reason to choose (cf. Thoma 2021). I assume that preferences are normative in the sense if $f \succ g$, then you should choose f in a pairwise choice between f and g.

[7] See Savage (1972 Chapters 3 and 4), as well as the helpful overview in Fishburn (1981). Fishburn (1989b), Machina and Schmeidler (1992), Grant (1995), Grant et al. (2000), Grant et al. (2008) each derive a subjective probability function using weaker assumptions than Savage's. Though Savage assumes an infinite set of states, many cases in this thesis involve a finite state space. Gul (1992) provides a Savage-style representation for finite sets of states by imposing a richness condition on the set of outcomes.

[8] Spohn (1977) notes that it is often useful to talk about theoretical terms in the absence of precise behavioural characterisations of them. Following Spohn then, I take it as methodologically fair to assume the existence of a subjective probability function and leave it for future work to discuss whether the theories discussed throughout this thesis can be given adequate representation theorems. Note that some cases I discuss could be described purely in terms of objective probabilities—we might follow Anscombe and Aumann (1963) in taking objective probabilities to be primitive and deriving subjective probabilities using objective ones as scaling devices.

outcomes—the *reduction principle* (see Fishburn 1988 p. 27). One of the issues that will arise in later chapters is whether the reduction principle holds or whether we miss something by ignoring states. So, while I assume that we can talk about an act's probability distribution over outcomes, I will not assume that this description exhausts everything we might care about.

With this setup in place, we are now in a position to characterise a range of dominance principles.[9]

0.2.1 Dominance and Preference

The first way that dominance-talk is used is when we compare acts. We say that one act dominates another when it is, in some sense, guaranteed to do better. But what does it mean to say that one act is guaranteed to do better than another?

0.2.1.1 State-wise Dominance

You are deciding where to go on your next holiday, Armidale or Tamworth. There are some things you know for certain, for example that Armidale has more expensive motels but that Tamworth has more country music. And then there are plenty of things you do not know, for example what the weather will be like and which town has better coffee. How should you decide on a holiday destination?

Your decision would be easier if you could resolve your uncertainty. If you could click your fingers and learn everything that matters to you—precise details about the weather, how good the coffee is, and so on—then you would *know* which option is better and simply prefer the better option.[10] But unfortunately few of us are omniscient, and we must compare options in light of our persistent uncertainty.

Enter State-wise Dominance. You need not know which state holds (i.e., how uncertainty is resolved) to recognise that one option might be better *in any case*. As is plausible, you might know that attending the Tamworth Country Music festival is such an unpleasant experience that,

[9] The Savage framework has been widely criticised, see Joyce (1999) and Bradley (2017). I adopt it here as a working framework primarily because it is a convenient language to state certain principles and problems in. Moreover, adopting a means-ends account of rationality seems to call for a formal distinction between means (acts) and ends (outcomes), so the Savage framework is natural (Buchak 2013 adopts the Savage framework in this spirit; cf. the comments in Fishburn 1981, Section 10).

[10] Or perhaps withhold forming a preference if you judge some outcomes to be incommensurable (cf. Hare 2010).

regardless of how the weather turns out or what the coffee is like, you prefer to be in Armidale over Tamworth. If you prefer one act regardless of which state holds, then you prefer it *simpliciter*. Formally, this principle is:

> **State-wise Dominance:** If for all $s \in S, f(s) \geq g(s)$, then $f \succcurlyeq g$. Moreover, if for some $s \in S, f(s) > g(s)$, then $f \succ g$.[11]

This principle strikes many as unimpeachable. 'Doing better however the world turns out to be' is a natural way of spelling out what it means for one act to be guaranteed to do better than another. You might of course think that State-wise Dominance applies in too few circumstances to count as a useful tool for simplifying real-world decisions—country music festivals aside, we rarely know that one act is better than another come what may. But even if we need more than State-wise Dominance, you might think that we cannot do with less.

Nonetheless, State-wise Dominance leads to counterintuitive verdicts in a range of cases. One problem is that our definition of a 'state' specifies that states hold *independently* of which act you choose. The (in)famous Newcomb's Problem, which I discuss in Chapters 2-4, shows that on plausible readings of independence, State-wise Dominance is far from obvious. Essentially, when your decision provides *evidence* about which state holds, some think that you can perform a state-wise dominated act if doing so is symptomatic of a propitious state holding. Moreover, most discussions of State-wise Dominance assume that preferences satisfy Transitivity. I argue in Chapter 5 that when this assumption fails, in particular when we allow preference cycles of the form $x > y > z > x$, then State-wise Dominance is again highly counterintuitive. So, it turns out that being 'better however the world turns out to be' may not be sufficient for being better *simpliciter*.

Setting aside what are arguably exotic cases, some question State-wise Dominance even in everyday cases. Consider the following (adapted from Diamond 1967):

> *Lolly*: David loves both Dora and Agnes and has a single indivisible lolly. He has no more reason to give the lolly to Dora than Agnes and no more reason to give the lolly to

[11] Note that we could adopt a weaker principle, which requires only that weakly prefer f to g if f does at least as well as g, regardless of which state holds:

> **Weak State-wise Dominance:** If for all $s \in S, f(s) \geq g(s)$, then $f \succcurlyeq g$.

We could further restrict the principle to apply to only to sufficiently likely states:

> **Non-Negligible State-wise Dominance:** Let $N_{f \succcurlyeq g} \subseteq S$ be the set of states such that $f(s) \geq g(s)$. If $C(N_{f \succcurlyeq g}) = 1$, then $f \succcurlyeq g$.

This lowers the bar for applying state-wise reasoning—so long as you assign no credence to states in which g beats f, f is at least as good as g.

> Agnes than Dora. He does not want to simply give the lolly to one of the women he loves, so he prefers to toss a coin (even at the expense of a dollar), thereby treating them fairly.

Since David is a simpleminded person, assume that there is no further uncertainty to resolve—once he knows who gets the lolly, everything else he cares about is settled. We can then model the case with two states, 'Coin Heads' and 'Coin Tails', and outcomes d (Dora gets the Lolly) and a (Agnes gets the lolly) such that $d = a$:

	Coin Heads	Coin Tails
Lolly to Dora	d	d
Lolly to Agnes	a	a
Toss Coin	$a - \$1$	$d - \$1$

Assuming David prefers more money to less, $a - \$1 < a$ and $d - \$1 < d = a$, David's preference for coin-tossing appears to violate State-wise Dominance. And yet David seems to be rational.

Some might therefore reject State-wise Dominance. Perhaps this case reveals that acts are more than the sum of their (state-wise) parts, meaning that the spread of outcomes induced by an act may matter in a way not reducible to the value of those individual outcomes.[12] Or, as many do (e.g., Broome 1991 Chapter 5; Buchak 2013 Chapter 4), you might think that this case shows how careful we must be in describing an agent's ends. If David really cares about fairness, then describing who gets the lolly *without* saying whether David has behaved fairly is an incomplete description of the outcomes he faces. This means that 'Dora gets the lolly' is not an outcome in the technical sense. The outcomes are rather more complicated propositions like 'Dora gets the lolly and is treated unfairly' and 'Dora gets the lolly and is treated fairly'.[13] So, many conclude, cases like David's demonstrate the care we must take in applying State-wise Dominance, not that the principle is false.

Even if David's case is not a counterexample to State-wise Dominance, it does emphasise something crucial about dominance principles. State-wise Dominance is only plausible when state- and outcome-descriptions are rich enough to capture everything you care about. The

[12] This is true of, say, some interpretations of *Mean-Risk Rules*, which evaluate acts based on both (i) the utility of the act in each state, and (ii) an act's overall 'risk-factor' (see, for example, Weirich 1984, p. 196).
[13] See Pettit (1991) and Broome (1991, Chapter 5) for discussions of how to individuate outcomes.

correct dominance principles therefore apply only in fully specified decision situations, not arbitrary descriptions of choices. (This, along with how rich state- and outcome- descriptions would have to be to capture *everything* we care about, is one of the reasons that Joyce 1999 and Bradley 2017 reject the Savage-framework altogether.)

0.2.1.2 The Sure-Thing Principle

Setting aside possible counterexamples to State-wise Dominance, let us ask whether it achieves everything we want from a dominance principle. Savage provides a classic example of dominance reasoning that is not merely an instance of State-wise Dominance:

> 'A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant to the attractiveness of the purchase. So, to clarify the matter for himself, he asks whether he would buy if he knew that the Republican candidate were going to win, and decides that he would do so. Similarly, he considers whether he would buy if he knew that the Democratic candidate were going to win, and again finds that he would do so. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say.' (Savage 1972, p. 21)

For the case to be minimally realistic, 'Republican wins' and 'Democrat wins' are not states—there is plenty left unresolved on finding out which party wins (which policies they implement, background economic conditions, etc.). Savage's Businessman therefore concludes that one act is better not because it is better however the *world* turns out to be, but because it is better conditional on anything he might *learn* (where this leaves some relevant facts unsettled). On the face of it, this seems like an unimpeachable piece of reasoning that we want our decision theory to respect—Savage (1972 p. 21) claims that he 'knows of no extralogical principle governing decisions that finds such ready acceptance'.

Savage translates the Businessman's reasoning into his framework using the following principle:

> **The Sure-Thing Principle (STP):** For acts f, f', g, g' and $E \subset S$ if (i) f and g agree on $\neg E$, (ii) f' and g' agree on $\neg E$, (iii) f and f' agree on E, and (iv) g and g' agree on E, then: $f \succ g$ if and only if $f' \succ g'$.

Though a bit of a mouthful, STP captures the idea that your preferences over acts should be governed only by what happens in states where those acts disagree. Your comparison between f and g goes the same way as your comparison between f' and g', since on states where f and g disagree, f is identical to f' and g is identical to g'. STP raises two related questions. The first is whether STP is true—must preferences be insensitive to states where acts agree? The second is whether STP really does capture the Businessman's reasoning—has Savage done the translation from informal reasoning to formal decision theory right?[14] Many answer 'no' to both questions.

Taking the first question first, STP appears to be overly restrictive in the *Allais Paradox* (Allais 1953). Consider the following two comparisons between bets on a fair, 100-ticket lottery:

	$1-89$	90	$91-100$
a	$\$1 \text{ million}$	$\$1 \text{ million}$	$\$1 \text{ million}$
b	$\$1 \text{ million}$	$\$0 \text{ million}$	$\$5 \text{ million}$

	$1-89$	90	$91-100$
c	$\$0 \text{ million}$	$\$1 \text{ million}$	$\$1 \text{ million}$
d	$\$0 \text{ million}$	$\$0 \text{ million}$	$\$5 \text{ million}$

Many prefer $a \succ b$ and $d \succ c$, though this violates STP. (Recall that STP says we can set aside states where acts agree—in each pair of gambles the bets agree on all but Tickets 90-100, and the two comparisons are identical on those states. So by STP, $a \succ b$ if and only if $c \succ d$.)

Though they violate STP, the Allais Preferences are easy to rationalise. Given a high chance of a good outcome on states where acts agree (such as when comparing a and b), we shy away from increasing the probability of bad outcomes, even if it means foregoing a chance of some fantastic good (e.g., winning $\$5 \text{ million}$). But if states where acts agree yield a bad outcome (such as when comparing c and d), we are prepared to bear more risk in order to secure a chance of that fantastic good. What happens on states where acts agree can play a role in how we assess risks borne on states where those acts disagree. So, contra Savage, it seems reasonable to play it safe in the first pair of gambles while taking a risk in the second.

0.2.1.3 Event-wise Dominance

Not so fast, replies the defender of STP! The discussion in the previous forces us to reject the apparently unimpeachable reasoning that drove Savage's Businessman (and recall that Savage

[14] Many will be familiar with the contrasting looks on students' faces on first hearing the story of Savage's Businessman and first seeing STP.

'knows of no extralogical principle' as plausible). Consider your preferences *conditional* on Tickets 90-100 being drawn—conditional on this event, you must either prefer $a \succcurlyeq b$ or $b \succcurlyeq a$. Say first that you prefer $a \succcurlyeq b$ conditional on Tickets 90-100 being drawn. Since c is identical to a and b is identical to d on these tickets, you must also prefer $c \succcurlyeq d$ on Tickets 90-100. And on Tickets 1-89 you weakly prefer $c \succcurlyeq d$ (indeed, you are indifferent since they yield the same outcome). So, you weakly prefer $c \succcurlyeq d$ conditional on Tickets 1-89 *and* conditional on Tickets 90-100. In either case, you weakly prefer $c \succcurlyeq d$, so Savage's Businessman says you should prefer $c \succcurlyeq d$ *simpliciter*. Now take the case that you prefer $b \succcurlyeq a$ conditional on Tickets 90-100. You weakly prefer $b \succcurlyeq a$ on Tickets 1-89 (since again they yield the same outcome on those tickets). So again, you weakly prefer $b \succcurlyeq a$ whether Tickets 1-89 or 90-100 were drawn, and Savage's Businessman says you should prefer $b \succcurlyeq a$ *simpliciter*. Regardless then of what you prefer on Tickets 90-100, you cannot simultaneously have the preferences $a \succ b$ and $d \succ c$, contra Allais.

Formally, we have just given an *Event-wise Dominance* argument against the Allais Preferences (Harsanyi 1977, for example, gives this kind of argument for STP). Let $E = \{E_1, \dots, E_n\}$ be a partition of S (i.e., for all $s \in S$, s is contained in exactly one of the E_i's). I will refer to each E_i as an *event*. Define $\succ_E$ as your preference *conditional* on event E holding, and let *Event-wise Dominance* be the principle that:

> **Event-wise Dominance:** If for all $E_i \in E$, $f \succcurlyeq_{E_i} g$, then $f \succcurlyeq g$. Moreover, if for some $E_i \in E$ $f \succ_{E_i} g$, then $f \succ g$.

This principle corresponds more closely to the intuitive reasoning in Savage's Businessman case than does STP. You subdivide your decision into simpler decisions by reasoning conditional on various events holding, and if you reach the same conclusion in each event, then you should reach that conclusion overall. Broome (1991, pp. 95-96) supports this reasoning by arguing that if you have more reason to choose an option regardless of which event holds, then you simply do have more reason to choose that option—after all, *one* of those events must hold. And, as we have seen, the Allais Preferences seemingly violate this principle (simply let $E = \{Tickets\ 1 - 89, Tickets\ 90 - 100\}$). Insofar as Event-wise Dominance captures a plausible coherence constraint between our conditional and unconditional views, *this* might be why we reject the rationality of the Allais Preferences.

Of course, the defender of the Allais Preferences might reject Event-wise Dominance for the same reason they reject STP: what happens in one event may influence the risks you are prepared to take conditional on another. Buchak (2013, p. 166) takes this line and claims that 'the local considerations about what happens in a particular state could remain the same in the presence of various outcomes in the other states—the desirability of each outcome could remain the same—but that these local considerations could be w*eighed* differently in the presence of other outcomes'. On Buchak's view, how you weigh risks in an event may depend on what happens in other events, contra Event-wise Dominance.

Is this response plausible? We might be left wondering precisely what is wrong with Savage's Businessman and note that it is unclear how we should weigh up the intuitive pull of Event-wise Dominance with the intuitive pull of the Allais Preferences. Might we not *tollens* in place of Buchak's *ponens* and conclude that since event-wise reasoning is overwhelmingly plausible, we should reject the rationality of interaction effects between events? We might also ask just how local considerations are '*weighed* differently in the presence of other outcomes'. Buchak has one story here (which I explore in Chapter 1), but is it the only plausible story, or can we tell one that better does justice to the intuitions driving Savage's Businessman?

Work from economics helps us to answer these questions. While a naïve appeal to Event-wise Dominance is incompatible with the Allais Preferences, more sophisticated versions of event-wise reasoning accommodate the Allais Preferences (and do justice to reasoning that drives Savage's Businessman).

To get at this point, note that Event-wise Dominance introduces a mysterious notion: *conditional preference*. What precisely is a conditional preference? As Grant et al. (2000, p. 185) note, conditional preferences are not directly revealed in choice—it is *preference*, not conditional preference that guides our decision-making. Of course, I might ask you what your preferences are supposing some fact, but then I have to rely on a verbal report and trust you that you are 'supposing' in the right way. Indeed, our ordinary concept of rationality might not dictate what it means to suppose E. So, we need to get clear on what precisely this bit of formalism '$\succ_E$' is.

Perhaps $\succ_E$ is the preference you would have if had you never entertained $\neg E$ possibilities (as if from birth you had been shielded from all talk of $\neg E$ possibilities)? Skiadas (1997a,b) notes, however, that the very lesson to draw from Allais might be that considering $\neg E$ affects how you assess some act given E. Skiadas claims that if learning $\neg E$ provides good news relative to E, then finding out E might induce feelings of regret—what he calls a 'subjective' component of the outcome, which is not captured in the 'objective' outcome-descriptions that we write in a decision table. Similarly, if learning $\neg E$ provides bad news relative to E, then finding out E

might induce feelings of elation (which again, is a subjective component of the outcome). This is a compelling idea, and I suggest on the right track. The problem is that Skiadas' approach is incompatible with a framework on which outcomes fully describe ends and so include everything you care about, whether objective or subjective.[15]

Alternatively, I suggest that what happens on $\neg E$ might influence your *risk-attitudes* conditional on E. If $\neg E$ is likely to yield some relatively good outcome, then you might be prepared to bear less risk conditional on E, since relatively good background conditions mean you prioritise avoiding bad outcomes. Similarly, if $\neg E$ is likely to yield some relatively bad outcome, then you might be prepared to bear more risk conditional on E, since relatively bad background conditions mean you are less concerned with avoiding bad outcomes. The lesson from Allais is that states where acts agree play a role in determining how you reason on states where those acts disagree.[16]

To illustrate, consider again the first pair of Allais gambles:

	$1-89$	$\mathbf{90}$	$\mathbf{91-100}$
a	\$1 million	**\$1 million**	**\$1 million**
b	\$1 million	**\$0 million**	**\$5 million**

Let's say you assess a and b conditional on Tickets 90-100 being drawn (in bold). You might think to yourself, 'I am very likely to get a million for sure so, conditional on 90-100 being drawn, I do not want to throw away a sure thing of a million by taking a risk on b', or put slightly differently 'Let me provisionally add to my stock of knowledge that one of Tickets 90-100 is drawn—I want to play it safe when I do so, since if in fact other tickets are drawn I am very likely to be a millionaire'. Supposing 90-100, the fact that 1-89 would have yielded a million means that you approach risks in a conservative way—you do not want to throw away a guaranteed boon. Compare this with the second pair of Allais gambles:

	$1-89$	$\mathbf{90}$	$\mathbf{91-100}$

[15] Sobel (1986, 1988a) rejects accounts of risk-sensitivity that appeal to facts about regret and rejoicing—outcomes are supposed to capture *everything* you care about! I do note, however, that Skiadas (1997b, p. 244) thinks it an advantage of his approach that outcome-descriptions may be partial rather than specifying everything that you care about, since outcomes are now the kind of thing we can feasibly describe. I am persuaded that the Allais Preferences are reasonable even if you attach no intrinsic value to subjective feelings of regret, elation, relief, and so on—this means that even if we accept the distinction between objective and subjective components of outcomes, we must look to more than the subjective components of outcomes to rationalise the Allais Preferences.

[16] Perhaps my approach and Skiadas' are compatible: you might take what happens on $\neg E$ to influence both subjective components of outcomes on E and your risk-attitudes supposing E.

c	$\$0\ million$	$\boldsymbol{\$1\ million}$	$\boldsymbol{\$1\ million}$
d	$\$0\ million$	$\boldsymbol{\$0\ million}$	$\boldsymbol{\$5\ million}$

Now let's say you assess c and d conditional on Tickets 90-100 being drawn (in bold). You might now think to yourself, 'I am very likely to get nothing for sure so, conditional on 90-100 being drawn, I am happy to take a chance on b', or 'Let me provisionally add to my stock of knowledge that one of Tickets 90-100 is drawn—I am happy to take a risk when I do so, since if in fact other tickets are drawn I walk away with nothing.'. Supposing 90-100, the fact that you are very likely to walk away with nothing means you approach risks in a more liberal way—there is no guaranteed boon to throw away. So, even if outcomes are complete descriptions of everything you care about, we need not equate 'your preferences conditional on E' with 'the preferences you would have had you only ever considered E possibilities'. We therefore ought not equate your conditional preference $\succ_E$ with the preference of some counterpart who has never faced the risks you face.

Perhaps then we should equate $\succ_E$ with the preference you will (or would) have on learning E. But what reason is there to equate your *current conditional* preferences with your *future unconditional* preferences? On learning E, your future self no longer faces $\neg E$ possibilities. But you do face $\neg E$ possibilities, which is a relevant difference between you and your future self. And again, the lesson from Allais might simply be that how you reason about risk depends on what you might, from your current epistemic perspective, get. So analysing $\succ_E$ as your future preferences on learning E is unmotivated.

In short, it is unclear why we should analyse $\succ_E$ in such a way as to screen off all information about what happens should E not hold—the risks you bear on $\neg E$ may affect your judgements supposing E. Formally, this means that conditional preferences need not be *separable*. $\succ_E$ is separable if for acts f, f', g, g' such that f, f' agree on E and g, g' agree on E, $f \succ_E g$ if and only if $f' \succ_E g'$. (Note how close this is to the STP!)[17] Though Savage (1972 p. 22) takes separability as something like an analytic truth, there is an important distinction between provisionally adding E to your stock of knowledge and assessing risks as if there are no $\neg E$ possibilities.

[17] See Skiadas (1997b, p. 244) for another definition of separability that captures a similar idea.

If we reject separability of conditional preferences, we can drive a wedge between Savage's Businessman and Allais. In Allais, we get violations of Event-wise Dominance *only if* we assume that conditional preferences are separable. But they need not be. And indeed, it turns out that Event-wise Dominance combined with non-separable preferences *is* compatible with the Allais Preferences. Skiadas (1997b, p. 252) calls this combination *Strict Coherence* and shows that Dekel's (1986) theory, which accommodates Allais Preferences, is compatible with Strict Coherence.[18]

Grant et al. (2000) provide another way of capturing the dominance reasoning that drives Savage's Businessman. They work entirely with unconditional preferences rather than introducing conditional preferences as does Skiadas. So, we need not introduce any new conceptual machinery—particularly conditional preferences that are not directly revealed by choice—in addition to preference. Grant et al. (pp. 172-175) then use Savage's Businessman's to motivate the axiom:

> **Decomposability:** For all acts f, g and any event E, if $f_E g \succ g$ and $f_{\neg E} g \succ g$, then $f \succ g$.

Say that f is 'Buy', g is 'Don't Buy', E is 'Republican Wins', and $\neg E$ is 'Democrat Wins'. Then Decomposability licenses the following Savage-style reasoning: 'Buying in the event that a Republican wins is an improvement over Not Buying. Similarly, buying in the event that a Democrat wins is an improvement over Not Buying. Since buying in either event is an improvement over not buying, buying is an improvement *simpliciter*.'

While Decomposability captures the Businessman's informal dominance reasoning in the Savage framework, it is again compatible with the Allais Preferences. Decomposability allows for a limited kind of non-separability between what happens in distinct events—it allows that $f_E g \succ g$ but that $g \succcurlyeq f$, though this is ruled out by STP. For example, if in Allais' case you have the preference $c_{1-89} d \succ d$, Decomposability permits that you prefer $d \succcurlyeq c$—what happens in the event of Tickets 1-89 can affect how you assess risks in the event that Tickets 90-100 are drawn.

And Decomposability is a non-trivial constraint. Indeed, for probabilistically sophisticated agents it turns out to be equivalent, given weak assumptions, to:

[18] See Skiadas (1997b, Proposition 3) and following for a discussion of Dekel's theory and its relation to non-separable preferences.

Betweenness: If $f \sim g$, then for all probabilities α, $f \sim \alpha f + (1-\alpha)g \sim g$. If $f \succ g$, then for all probabilities α, $f \succcurlyeq \alpha f + (1-\alpha)g \succ g$.[19]

Betweenness rules out premiums for (or against) pure randomisation: when you are indifferent between two options, you cannot make them better or worse by tossing a coin to decide between them. A remarkably intuitive constraint—imagine how difficult life would be if we could not toss coins to make up our minds! But Betweenness is again compatible with the Allais preferences—I discuss an Allais-accommodating, Betweenness-satisfying theory in Chapter 1.

Betweenness and Decomposability *are*, however, incompatible with some popular decision theories designed to accommodate the Allais Preferences. For example, *rank dependent* models like Quiggin's (1982) *Rank-Dependent Utility Theory* and Buchak's (2013) *Risk Weighted Expected Utility Theory* violate Betweenness.[20] So Decomposability, a natural way of vindicating Savage's Businessman without ruling out the Allais Preferences, does rule out many popular accounts of the Allais Preferences.

This raises an interesting question. Should we accept rank-dependent theories of rationality or uphold some form of event-wise reasoning along the lines of Decomposability? For Buchak, the answer is presumably that we should reject event-wise reasoning altogether.[21] But we need not go this route. Given the *prima facie* plausibility of Betweenness and the intuitive appeal of Savage's Businessman, we should ask whether Decomposability- and Betweenness-satisfying theories are normatively defensible, and we should assess the virtues of Betweenness-satisfying theories to get a holistic sense of which dominance principles our best theory of instrumental rationality upholds. In Chapter 1, I argue that Betweenness-satisfying theories are indeed normatively defensible and have independent advantages over Betweenness-violating ones.

Before moving on, note that I have vindicated one of my initial claims: while the basic idea behind dominance is straightforward enough, it is challenging to spell out in precise detail. There are multiple ways of getting at the intuition behind Savage's Businessman, and our decision as to how and whether to formalise that intuition has substantive implications for which preferences are rational.

[19] See Grant et al. (2000, Proposition 3). The weak assumption are: Savage's P1 (preferences are complete and transitive), P3 (Eventwise Monotonicity, which I discuss in the next section), and P6 (Event Continuity, which is a richness condition on the state space).

[20] These theories satisfy Betweenness only in the case that they adopt a risk-neutral probability weighting function, in which case they cannot accommodate the Allais Preferences (see Camerer 1989, pp. 77-78).

[21] Perhaps slightly more accurate would be to say that Buchak would reject that event-wise reasoning is rational *in all cases*. Buchak (2013, pp. 107-110) accepts STP applied to a special class of cases, those in which f, f', g, g' are *comonotonic* (i.e., they induce the same weak preference ordering on states).

0.2.1.4 Mixture Dominance

Before moving on, it is worth mentioning a slightly strengthened version of Betweenness, Fishburn's (1982):

> **Mixture Dominance:** If $f \succ g, f \succ h$, then for all probabilities $\alpha, f \succ \alpha g + (1-\alpha)h$. If $f \sim g, f \sim h$, then for all probabilities $\alpha, f \sim \alpha g + (1-\alpha)h$. If $f \prec g$, $f \prec h$, then for all probabilities $\alpha, f \prec \alpha g + (1-\alpha)h$.[22]

This says that if one act is better (worse) than two others, then it is better (worse) than any gamble between them. Again, this corresponds to a natural dominance-like thought: a gamble between options that are worse (better) than some alternative is in no sense as good as (worse) than that alternative.

Mixture Dominance is closely related to Betweenness, indeed it is equivalent given Transitivity.[23] So, Betweenness-violating theories like Buchak's also violate Mixture Dominance. In the absence of Transitivity, Mixture Dominance will serve as the backbone of a remarkably elegant rule for comparing risky acts, Fishburn's (1982) *Skew-Symmetric Bilinear Utility Theory*, which I discuss in Chapter 5.

0.2.1.5 Stochastic Monotonicity and Stochastic Dominance

The Sure-Thing Principle fails, so many say, and there are situations where 'substituting in' a preferable act does not improve things: $f \succ g$ does not entail $f_E h \succ g_E h$. But paradigmatic cases where this reasoning fails involve *risky acts* (f,g, h are not degenerate). If we improve an act by substituting out a single *outcome* for a better one, then surely this results in an improvement. This idea is widely accepted. Let E^* be the set events to which you assign non-zero probability, then (Savage 1972, P3):

[22] Fishburn simply calls this principle Dominance—for obvious reasons, that term would be ambiguous in the present context!

[23] Proof: Say first that Mixture Dominance holds. If $f \sim g$, we know $f \sim f$ trivially and so by Mixture Dominance $f \sim \alpha f + (1-\alpha)g \sim g$, giving us Betweennness. Now say that Betweenness holds. If $f \succ g$ and $f \succ h$, then by Transitivity either $f \succ g \succ h$ or $f \succ h \succ g$. By Betweenness, in the first case $g \succ \alpha g + (1-\alpha)h \succ h$, and in the second $h \succ \alpha g + (1-\alpha)h \succ g$. In either case by Transitivity $f \succ \alpha g + (1-\alpha)h$. This gives us the strict preference part of Mixture Dominance (the proof for the other parts is the same).

Event-wise Monotonicity: For any act f, $E \in E^*$, and $x, y \in O$, $x > y$ if and only if $x_E f \succ y_E g$.

Event-wise Monotonicity is entailed by the following principle, formulated in terms of probability distributions over outcomes (see Grant et al. 2008, p. 376):

Stochastic Monotonicity: For any act f, probability $\alpha > 0$, and $x, y \in O$, $x \geq y$ if and only if $\alpha f + (1-\alpha)x \succcurlyeq \alpha f + (1-\alpha)y$.[24]

Stochastic Monotonicity in turn is entailed by *First-Order Stochastic Dominance*. Take two acts $f = \alpha_1 o_1 + \cdots \alpha_n o_n$ and $g = \beta_1 o_1 + \cdots + \beta_n o_n$ (note that in order to describe f and g as inducing probability distributions over the same set of outcomes, some α_i and β_i might be zero):

First-Order Stochastic Dominance: If for all outcomes o, $\sum_{i:o_i \leq o} \alpha_i \leq \sum_{i:o_i \leq o} \beta_i$ and for some o, $\sum_{i:o_i \leq o} \alpha_i < \sum_{i:o_i \leq o} \beta_i$, then $f \succ g$.

First-Order Stochastic Dominance again captures an extremely natural thought: if for each outcome f is as likely as g to bring about something better than that outcome (and for some outcome *more* likely to bring about something better than that outcome), then f is preferable to g.[25]

Stochastic Dominance is widely discussed—it lands in the goldilocks zone of substantive enough to have serious implications for decision theory while being highly intuitive. Many take it to a be a non-negotiable feature of a plausible decision rule (e.g, Tarnsey 2020, Wilkinson Forthcoming)—Meacham (2020 p. 1000) goes as far as to say that it 'seems like a Moorean fact'. And it is easy to see the intuitive pull: if a standard of rationality says that you do not make things better by increasing the probability of the good, then it is unclear what incentive you have to follow that standard of rationality.

And the implications of Stochastic Dominance are indeed substantive. Tarnsey (2020 Section 5.4) proves that given appropriate conditions of background uncertainty, agents with transitive

[24] See Grant et al. for discussion of this principle in relation to other monotonicity conditions in the Savage framework. This principle is elsewhere (e.g., Grant 1995) called the 'Axiom of Degenerate Independence' since it is essentially von Neumann and Morgenstern's Independence restricted to degenerate acts. Event-wise Monotonicity entails Stochastic Monotonicity when the underlying state space is sufficiently rich (see Machina and Schmeidler 1992, p. 60).

[25] Note that there are also various kinds of *Higher-Order* (or *n^{th}-order*) Stochastic Dominance. That you prefer f to g when f n$^{\text{th}}$-order stochastically dominates g is not a purported rational constraint. Rather, respecting n$^{\text{th}}$-order stochastic dominance is typically taken to characterise some type of preference. For example, that you respect Second-Order Stochastic Dominance is often taken as a definition of *risk-aversion* (I return to this point in Chapter 1).

preferences who obey First-Order Stochastic Dominance will approximate Expected Utility Maximisers in their choices. (Here 'appropriate conditions' means a heavy-tailed probability distribution over outcomes—roughly that your credence in more extreme outcomes decreases no more quickly than as in a *Laplacian* distribution—see Tarsney p. 12 for details). He also notes that Stochastic Dominance can be a useful tool for comparing gambles with infinite values (e.g., betting on heaven in Pascal's Wager). This leads Tarsney to suggest that Stochastic Dominance may indeed be *all* we need from a decision theory. Hedden (2020) puts Stochastic Dominance to work to solve collective action problems, even those in which affected peoples' well-beings might not be commensurable. And economists are particularly interested in Stochastic Dominance as it allows us to specify situations in which rational agents with different utility functions will nonetheless make the same choices.[26] For something like a Moorean fact, we get a lot out of Stochastic Dominance.

There are, however, at least two situations in which Stochastic Dominance might fail. The first is due to Hare (2010) and involves dropping the assumption that preferences are *Complete*. I say that you have a preference gap between two outcomes if you do not weakly prefer one to the other. Take two such outcomes, r (perhaps 'Reading David Copperfield') and s (perhaps 'Seeing the magician David Copperfield'). Assume also that your preference gap is insensitive to mild sweetening: though $r + \$1 \succ r$ and $s + \$1 \succ s$, you have no preference between either $r + \$1$ and s or $s + \$1$ and r (you are not fickle enough to let a dollar break your indecision between two great works of art).

Now consider the following gambles on the toss of a fair coin (the coin was tossed yesterday, so nothing you do will influence the result of the coin-toss):

	Coin Heads	Coin Tails
Plain Gamble	r	s
Sweetened Gamble	$s + \$1$	$r + \$1$

The Sweetened Gamble ought to be preferred by the lights of Stochastic Dominance. And yet there seems to be a compelling argument against preferring the Sweetened Gamble: *however* the coin landed, you do not prefer the Sweetened Gamble. Indeed, something close to State-wise Dominance requires us to reason in this way:

[26] See Fishburn (1989a), which surveys various surveys.

Negative State-wise Dominance: If for all $s \in S$, $\neg(f(s) > g(s))$, then $\neg(f \succ g)$.

This negative principle has much of the appeal of State-wise Dominance itself—if there is no way the world could be such that some act is better, then that act is not better *simpliciter*. Schoenfield (2014) argues that anyone who prefers the Sweetened Gamble is guilty of 'expected value fetishism'—letting facts about what is instrumentally valuable outstrip facts about actual value, which is what you ultimately care about. Buchak (2013, p. 75) similarly endorses a 'Reasons for Betterness' principle that leads to Negative State-wise Dominance. So, we might reject Stochastic Dominance as a general principle.

Others (e.g., Bader 2018 and Wilkinson 2020) uphold Stochastic Dominance in light of Hare's case. Bader, for example, argues that concern for promoting value (or, presumably, a commitment to means-end rationality) does not require reasons to come from intra-state comparisons. If you prefer the Sweetened Gamble, this may still be because of value considerations, even if not 'actual value' in Schoenfield's sense. Bader claims (p. 505) that when comparing acts, we should not be concerned with how things go in each state but 'how good things are relative to what could just as well have come about'. So, Bader reasons in Hare's case that *ex ante* the Sweetened Gamble *could* bring about $r + \$1$, while the Plain Gamble *could just as well* bring about r. Similarly, *ex ante* the Sweetened gamble *could* bring about $s + \$1$ while the Plain Gamble *could just as well* bring about s. Of course, there is no way the world could be such that Sweetened Gamble brings about $r + \$1$ *and* Plain Gamble brings about r. But, on Bader's view, such intra-state comparisons are irrelevant. What matters is that *ex ante* for any outcome the Unsweetened Gamble might yield, the Sweetened Gamble could just as well bring about a better outcome. That makes it a better means to your ends.

We now have two pictures of rationality emerging, one that evaluates acts based on intra-state comparisons and another that evaluates acts based on which outcomes they *might* bring about, regardless of which outcomes occur in which states. Bader opts for the latter and so upholds Stochastic Dominance. Schoenfield opts for the former and upholds Negative State-wise Dominance.[27] Interestingly, two seemingly unimpeachable dominance principles pull in different directions. This will become particularly important when I discuss preference cycles in Chapters 5 and 6.

Stochastic Dominance has also been challenged in cases involving small probabilities. Some argue that an outcome with low enough probability, a so-called *de minimis* probability, contributes

[27] Doody (2019a, 2019b) notes structural challenges facing both responses.

nothing to an act's instrumental value.[28] Let such a probability be δ. The defender of *de minimis* probabilities says that the following two acts should be evaluated equivalently:

$$f = \$0$$

$$g = (1-\delta)\$0 + \delta\$1{,}000{,}000$$

The δ-increase in the probability of a million here contributes nothing. Some take this violation of Stochastic Dominance as a reason to reject *de minimis* probabilities (e.g., Isaacs 2016, Lundgren and Stefánsson 2020, Wilkinson Forthcoming). Others think that *de minimis* probabilities are required by considerations from philosophy of language (Smith 2016)[29] or are permitted by our best theory of rationality (Buchak 2013, pp. 73-74)[30]. And *de minimis* probabilities allow us to avoid notorious paradoxes that arise when acts assign non-zero probability to an infinite number of outcomes (e.g., the *St Petersburg Paradox*, see Peterson 2019, and the *Pasadena Paradox*, see Hájek and Nover 2004). And defenders of *de minimis* probabilities can retreat to:

> **Weak Stochastic Dominance:** If for all outcomes o, $\sum_{i:o_i \preccurlyeq o} p_i \geq \sum_{i:o_i \preccurlyeq o} q_i$, then $f \succcurlyeq g$.

This principle says that if one act brings about as good a chance of the good as another, it does *no worse*. So, the defender of *de minimis* probabilities may maintain that some increases in the probability of the good are rationally negligible, making things neither better nor worse. Indeed, if we take the intuition behind such probabilities seriously, this may be the right thing to say (though see Lundgren and Stefánsson 2020, Section 5 for a response).

A final point worth noting is that Stochastic Dominance needs to be applied with caution in cases of *multivariate* goods—outcomes that are themselves represented as bundles of distinct goods. For example, say that you care about xylophones and yachts, combinations of which are represented as ordered pairs of the form (x, y), representing the number of xylophones and yachts you receive, respectively. It has been widely noted (e.g., Scarsini 1988; Grant et al. 1992) that having preferences that satisfy Stochastic Dominance over individual goods does not guarantee Stochastic Dominance over multivariate outcomes. For example, say that you have the following preferences when gambling for individual goods:

[28] This was Buffon's response to the well-known St Petersburg game (see Peterson 2019, Section 5).

[29] Smith argues that language is not infinitely precise and so norms that mention precise probabilities (e.g., 'ignore events of probability zero') should be tolerant (e.g., 'ignore events whose probability is $0 + \delta$').

[30] On Buchak's view, agents weight outcomes not by probabilities but *risk-weighted* probabilities. Some risk-weighting functions may assign zero weight to outcomes with non-zero probabilities.

$$\frac{1}{2}(5x)+\frac{1}{2}(0) \succ 2x$$

$$\frac{1}{2}(5y)+\frac{1}{2}(0) \succ 2y$$

At the same time, there are interaction effects within bundles of good: you would rather two xylophones and two yachts to one of one and five of the other (what good are yachts without music?):

$$(2,2) > (0,5) \sim (5,0)$$

Now we have an apparent tension. You might reason to yourself, 'I disprefer 2 x's to a coin toss between 0 and 5 x's' and 'I disprefer 2 y's to a coin toss between 0 and 5 y's' and conclude 'I therefore disprefer 2 x's *and* 2 y's to a coin toss that yields both 0-or-5 x's and 0-or-5 y's'. But this amounts to the preference:

$$(2,2) \prec \frac{1}{2}(0,5)+\frac{1}{2}(5,0)$$

And this violates Stochastic Dominance: the option on the right is *guaranteed* to bring about a worse (multivariate) outcome than the one on the left.

Some economists think that Stochastic Dominance 'does not possess the same normative strength' as a principle for multivariate goods as it does univariate ones (e.g., Machina and Schmeidler 1992, footnote 17). And certainly, if people reason about individual goods, then as a descriptive matter we might expect them to violate Stochastic Dominance for multivariate outcomes. From a normative perspective, however, the right response seems to be that if outcomes really are multivariate, then you ought not reason about those outcomes simply in terms of their component parts. Of course, by choosing $(2,2)$ over $\frac{1}{2}(0,5)+\frac{1}{2}(5,0)$ you are violating your own preferences over individual *goods*: you choose 2 x's over $\frac{1}{2}(5x)+\frac{1}{2}(0)$. But note that 'receiving 2 x's' is not itself an *outcome*—a complete description of everything you care about specifies how many x's and how many y's you get. So again, multi-variate cases highlight the reasonable dominance principles apply only to fully described decision situations, not to incomplete descriptions of the choices you face.

(A brief aside: if there are independent normative constraints on multivariate outcome-evaluations, then we may indeed get a genuine tension between Stochastic Dominance and preferences over individual goods. Say that outcomes (x,y) represent the wellbeing levels of two individuals. Both individuals prefer $\frac{1}{2}(5) + \frac{1}{2}(0) \succ 2$. You, however, are inequality averse and so have preferences $(2,2) > (5,0) \sim (0,5)$. But you are also convinced on moral grounds that a social choice rule should respect unanimous preferences—if everyone prefers some lottery *ex ante*, then you should prefer that lottery as the social planner. Since everyone prefers to get a coin toss on 0-or-5 rather than 2 units outright, *your* preference over risky social outcomes must therefore be $(2,2) \prec \frac{1}{2}(0,5) + \frac{1}{2}(5,0)$, which as before violates Stochastic Dominance. In response, you might deny that respecting unanimous *ex ante* preferences is a reasonable constraint when it comes to social choice. Or, you might think that a commitment to certain decision procedures (e.g., letting the majority decide) means that Stochastic Dominance must go. Or, you might think that certain combinations of risk-aversion and inequality-aversion are incoherent.)

0.2.2 Dominance and Choice

So much for preferences. I now turn to a second use of dominance-talk, as applied to *choices*.[31] When confronted with a set of options, in what sense can once choice from that set be better than another in every sense?

Davidson et al. (1955, p. 145) introduce the following principle:

> **Non-Dominated Choice:** A rational choice is one which selects an alternative to which none is preferred.[32]

Formally, let $\boldsymbol{c}(\cdot)$ be a *choice function*, which maps each set of options to a subset of permissible options. We can then restate Davidson et al.'s principle, that for option set A:

> **Non-Dominated Choice:** If $f \in \boldsymbol{c}(A)$, there is no $g \in A$ such that $g \succ f$.

Note that when we say that f is 'dominated' in the set A because $g \succ f$, it need not be that g state-, event-, stochastically-, or otherwise dominates f. This might make us question whether the Non-Dominated Choice Principle really fits the mould of a *dominance* principle. I think it does

[31] In what follows I take preferences as primitive and then ask what it is permissible to choose in light of those underlying preferences. That I take preferences, rather than choice functions, as primitive is a substantive assumption throughout (see the discussion in Hansson and Grüne-Yanoff 2017, Section 5.3).

[32] Davidson et al. do not use the word 'dominance'—I take that terminology from Gustafsson (2013, p.460; 2016, p. 62). Throughout I will refer to this principle as Davidson et al.'s Non-Dominated Choice constraint, though strictly speaking it is the constraint, not the name, that comes from Davidson et al.

in the following sense: once you have formed a preference $g \succ f$, then *from the perspective of choice* there are no further facts that could make you say that f is as good a means to an end as g. Therefore, choosing f is in no sense as good as choosing g. So, while we must note the shift in our use of the word dominance, Davidson et al.'s principle captures another reasonable sense in which one act can be unambiguously better than some alternative.

The Non-Dominated Choice Principle can be challenged. One challenge involves infinite option sets. For example, say that I offer you any integer number of dollars that you like (I assume that your preferences are strictly increasing in money). You can select any option from:

$$A = \{\$1, \$2, \$3, \dots\}$$

Since for any choice of $\$y$, there is some $\$x$ such that $\$x \succ \$y$, the Non-Dominated Choice Principle deems *anything* you choose here impermissible. But surely the correct theory of rationality cannot give rise to rational dilemmas—there is always some permissible thing that you can do.[33] Moreover, it is hard to fault your preferences here—the world is just structured in such a way that there is no 'best' thing. Perhaps the Non-Dominated Choice Principle then is too strong.[34]

The Non-Dominated Choice constraint also creates tensions for agents with *preference cycles*. Say that we drop Transitivity and allow that you have preferences $a \succ b \succ c \succ a$. Consider the option set:

$$B = \{a, b, c\}$$

Whatever you choose from B strictly dispreferred to an available alternative, so you are forced to make an irrational choice by the lights of the Non-Dominated Choice Principle. Again, perhaps again the Non-Dominated Choice Principle is too strong. Or, as Levi (2002) and Gustafsson (2013) argue, this may be an argument against the rationality of cyclic preferences (I return to this issue in Chapter 5).

We might amend the Non-Dominated Choice Principe to accommodate infinite and non-transitive option sets. For example, we could say that for any option set A:

[33] We do not need an unlimited supply of money to get this problem—even *I* can offer you any dollar amount you like from the open interval $(0,1)$.

[34] See Arntzenius et al. (2004) and Bartha et al. (2014) for further problem cases that arise when option sets are infinite.

Weak Non-Dominated Choice: If $f \in \boldsymbol{c}(A)$, then for any $g \in A$ such that $g \succ f$ there is some $h \in A$ such that $h \succ g$.

This is not quite as demanding as Davidson et al.'s principle. It says that dominated options are impermissible *provided* that each dominating option is not itself dominated. It is still a non-trivial constraint since if a set contains a maximal element (one that is not strictly dispreferred to another), then it prohibits your choosing any non-maximal element. All finite option sets have a maximal element, as do some infinite sets (e.g., $A' = \{\$0, -\$1, -\$2, \dots\}$) and some sets containing cycles (e.g., $B' = \{a, b, c, d\}$ with $a \succ b \succ c \succ a$ but $d \succ a$, $d \succ b$, and $d \succ c$). Indeed, by definition for each option set either (i) there is some option that is weakly preferred to all others, or (ii) every option is strictly dispreferred to another. So, the Weak Non-Dominated Choice Principle always permits something. We might therefore move from the Non-Dominated Choice to some more permissive constraint along the lines of Weak Non-Dominated Choice. I will go further in Chapter 6 and argue that we might reject both.

So again, getting the details of dominance right matters. The Weak Non-Dominated Choice constraint illustrates a tension that will arise in Chapters 5 and 6—should we impose constraints on rational preferences so that we respect some constraint(s) on choice, or should we revise choice constraints in light of seemingly reasonable preferences?

Are there other dominance-like constraints on choice functions? Here are a couple of candidates from Sen (1971, p. 314):

$\boldsymbol{\alpha}^*$**:** If for option sets A, A', if $x \in \boldsymbol{c}(A \cup A')$ then $x \in \boldsymbol{c}(A) \cup \boldsymbol{c}(A')$

$\boldsymbol{\gamma}^*$**:** If for option sets A, A', $x \in \boldsymbol{c}(A)$ and $x \in \boldsymbol{c}(A')$, then $x \in \boldsymbol{c}(A \cup A')$.

These are often referred to as *contraction* and *expansion* consistency principles, but it does not take much to see them as analogous to dominance principles already discussed.[35] For example, we might think of $\boldsymbol{\gamma}^*$ as a 'Subset-wise Dominance' principle: rather than reasoning that one option does better in each *state* or *event*, you reason that one option is choiceworthy in each *subset* of available options, so it is choiceworthy *simpliciter*. Of course, a choice from $A \cup A'$ is not literally a choice from one of the subsets A or A'—you choose directly from $A \cup A'$. But heuristically, we might think of choosing from an option set as involving first a choice of subset, then a direct choice from that subset. And, so the thought goes, if you have most reason to choose something

[35] Steele (2010, pp. 464-465) makes this connection, though implicitly via von Neumann & Morgenstern's Independence Axiom, which is analogous to Savage's STP.

in each subset, then you must choose it overall. Subsets here play the role of events for Savage's Businessman.

And just as interactions between events threatened (or at least complicated) Savage's event-wise reasoning, they threaten (or complicate) subset-wise reasoning. Consider the following Sen-inspired case:

> When choosing between a *small* and *medium* slice of cake, you should choose the small slice on the grounds that not doing so would be rude. When choosing between a *small* and *large* slice of cake, you should choose the small slice on the grounds that not doing so would be rude. But when choosing between *small*, *medium*, *and large* you should not choose the small slice. You prefer more cake to less, and it is not rude to choose medium when large is available!

In one subset $\{small, medium\}$, politeness requires $small$, and the same goes for the other subset $\{small, large\}$. But your reasons against taking the medium slice in the first subset depend on $large$ not being present. So, $medium$ has a property locally (i.e., in the set $\{small, medium\}$) that it does not have globally (i.e., in the set $\{small, medium, large\}$). Just as Allais drove a wedge between global and local considerations in cases of risk, Sen drives a wedge between global and local considerations when choosing from option sets.

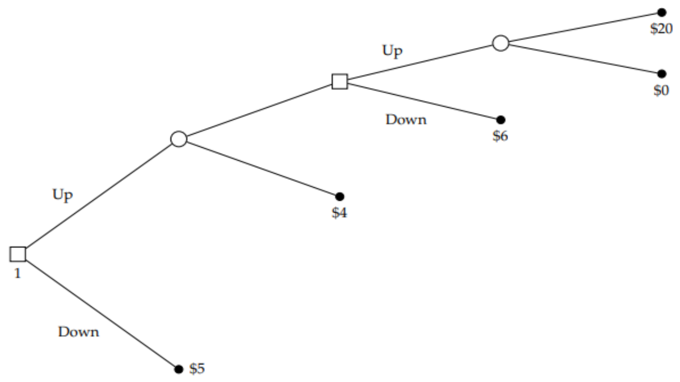
The discussion over Sen's principles is extensive. Just as some re-individuate outcomes in cases of risk (e.g., Broome 1991), you might re-individuate outcomes so that options can depend on what else is available. For example, maybe 'Choosing medium when large is available' is not the same as 'Choosing medium when large is unavailable'—the former has a normatively significant property (politeness) that the latter lacks. Or, just as some reject various dominance principles, we might reject the above constraints as principles of rationality. Here I can do no more than to emphasise the analogy between choice constraints and other dominance principles.

0.2.3 Dominance, Sequences, and Plans

So much for dominance constraints on preferences and choices. I now move to one final way that dominance-talk gets used—as applied to *plans* and *sequences* of options.

We do not just make one-off choices. Rather, we engage in courses of action over time that involve many distinct choices. Moreover, we make contingency plans for what to do in various events—a *plan*. What then does it mean to say that one plan is guaranteed to do better than another?

Tree-diagrams provide a convenient way of thinking about sequential decision problems. Each choice you face is represented by a square, and final holdings, which are outcomes, are represented by bold dots. Sometimes nature makes a choice, and nature's choices are represented by circles. Formally, a decision tree is a directed graph with a single root node, every arrow pointing away from that root node,[36] and every path terminating with a node that corresponds to an outcome. Nodes come in two kinds—choice nodes (which correspond to things you can choose) and chance nodes (which correspond to things that nature chooses). For example:



The above graph represents the situation where you make initially make a choice. If you choose to go 'Up' at 1, nature makes a choice (say, a coin is tossed), that yields either $4 or a further choice between $6 outright or some other lottery, say another coin toss, over $0 and $20; if you choose 'Down' at node 1, you get $5 outright. This way of representing decisions is often called *extensive form*.[37]

A *sequence* is any path through the decision tree, and a *plan* specifies for each contingency (each move that nature might play), which choice you make in response to that contingency. We can also represent plans in *normal* form by listing out the final payoffs that each plan might yield, depending on the moves nature might play. For example, in normal form the plans in the above diagram are:

[36] Indeed, when I draw decision trees I do not include the 'arrows', since we always move from left to right.

[37] I assume that agents are always aware of the decision tree they are in (i.e., what possible moves nature might play, what possible choices they might face, and what the outcomes are).

Plans	*Payoffs*
Down at 1	$5
Up at 1, up at 2 if Heads	$10 if Heads, $4 if Tails
Up at 1, down at 2 if Heads	$10 if Heads, $0 if Tails-Heads, $20 if Tails-Tails

Note that the column on the right does not mention times or order of moves, just states of nature and what each plan yields in that state. We can think of normal form descriptions of plans as abstracting away from the dynamic nature of plans and describing them as Savage-acts. (Importantly, on many theories of options,[38] normal form plans are not options since there is no time at which you can bring them about at will. By abstracting away from the dynamic nature of your decision, normal form presentations describe plans *as if* they were options.)

Hammond (1977, 1988) argues that optimal plans in extensive form (i.e., the plans you choose considering each choice you make at each time) should correspond to optimal plans in normal form (i.e., the plans you would choose when considering only probability distributions over final outcomes). This condition is known as *normal-extensive form coincidence* (NEC). If we think of plans in normal form as idealised options, then this is a dynamic analogue to Davidson et al.'s Non-Dominated Choice constraint: you have no reason to take an (idealised) option that is strictly dispreferred to another, so extensive form solutions ought to be normal form solutions as well.

NEC strikes some as a plausible constraint, though others question its motivation (e.g., Seidenfeld 1994, Levi 2002, Steele 2010). But again, dominance to the rescue! Steele (2010, Section 4) considers rejecting NEC but maintains that a plan is irrational if, compared to some alternative plan, it yields a worse outcome in *every* state of the world. So, Steele adopts a stronger reading of what it takes for a plan to be dominated: not just that it is dispreferred to some other normal form plan, but that it yields worse outcomes regardless of which moves nature plays. Steele (p. 470) therefore endorses:

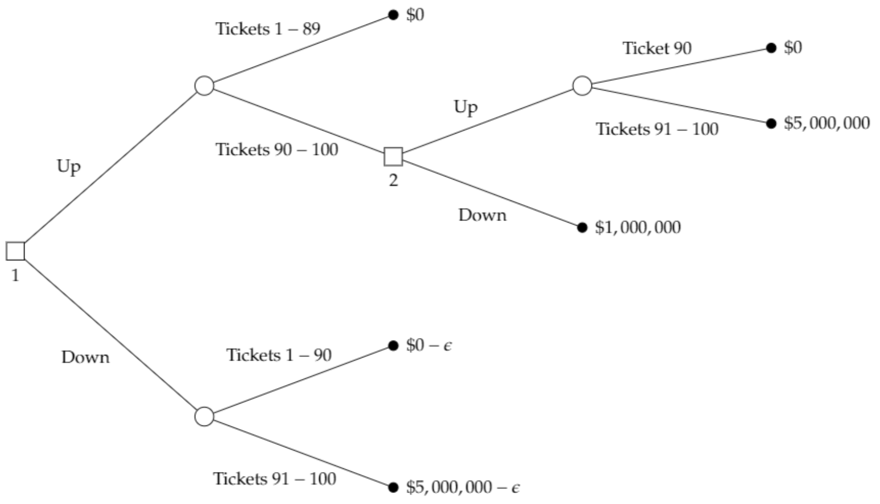
> **Non-Dominated Plans:** A rational plan is one that is not dominated (in every state-by-state comparison) by some available alternative plan.[39]

Non-Dominated Plans places substantial constraints on what preferences are rational—for a plan to be rational there must be *some* way the world can be such that you end up with a

[38] See for example Jeffrey (1983), Weirich (1983), and Hedden (2012).

[39] Steele (2010) provides a detailed discussion of the implications of moving from NEC to Non-Dominated Sequences.

preferable outcome. Consider the following dynamic version of Allais, in which you can *either* take d outright for a small fee, $\$\epsilon$, or make your decision after learning whether Tickets 1-89 or Tickets 90-100 were drawn:



Say that you have the Allais Preferences and assume that you prefer a million for sure over a lottery yielding $.9$ probability of five million and a $.1$ probability of nothing.[40] Then if you find yourself at node 2, you choose 'Down'. That means by choosing 'Up' at node 1, you effectively guarantee yourself the lottery $.89(\$0) + .11(\$1)$, which is c from the original Allais case! And since you disprefer $c \prec d$, you are prepared to pay a small fee, $\$\epsilon$, to get d over c. So, by choosing 'Down' at node 1 you satisfy your preferences and get $d - \$\epsilon$ over of c. But this plan is dominated: if you were to choose 'Up, Down', you would effectively get d outright and end up $\$\epsilon$ richer, whichever ticket is drawn.

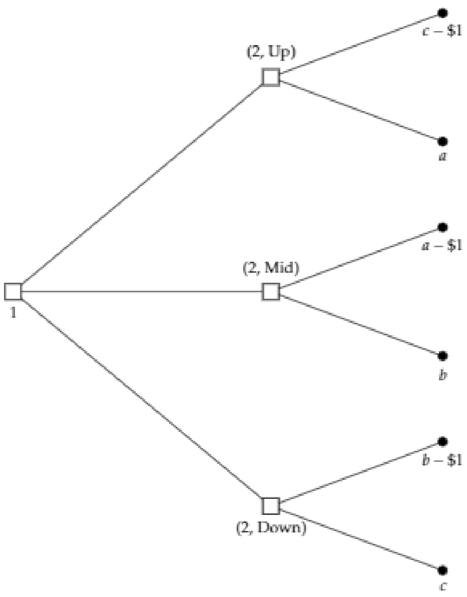
There have been a range of responses to this. Hammond (1988) rejects the rationality of the Allais Preferences. Machina (1989), however, denies that an otherwise rational agent with the Allais Preferences really does choose dominated plans. Machina denies *dynamic separability*. The demonstration of sure loss above assumed that at choice node 2 you choose 'Down' rather than 'Up'. But this assumes that you treat node 2 as a decision separate from the risks you have previously borne. Machina thinks that you may indeed choose the 'Down' option at choice node 2 since (p. 1645) 'an agent with non-expected utility/non-separable preferences feels risk which is borne but not realized is gone in the sense of having been *consumed* (or "borne"), rather than

[40] The parallel argument in the case that you prefer the lottery to the million is given in Buchak (2013, pp. 184-187).

gone in the sense of *irrelevant*'. The risks you bore in getting to node 2 might affect your choice at node 2.[41] And if so, the 'Up, Up' plan might be available to you. Others, for example Levi (2002) and Hedden (2015), deny that the dominating plans here are available as things you can do in a normatively significant sense (more on this shortly).

The Non-Dominated Plans constraint also puts pressure on the rationality of cyclic preferences. Consider an arbitrary preference cycle $a > b > c > a$ and the following scenario from Cantwell (2003):

> You can select any one of $\{a, b, c\}$. After you make your decision, I will offer you one more choice. If you chose a, I offer you c for a small fee (say, $\$1$). If you chose b, I offer you a for a small fee (say, $\$1$). If you chose c, I offer you b for a small fee (say, $\$1$).[42] In extensive form:



Here, there are plans that lead to a, b, and c—those in which you choose 'Down' at any of the second choice nodes. And there are plans that lead to $a - \$1$, $b - \$1$, and $c - \$1$—those in which you choose 'Up' at any of the second choice nodes. And you know that you will choose

[41] This is often defended on the grounds that past risks or decisions might influence your subjective experience of the outcomes you receive (see Loomes and Sugden 1986 and Machina 1989, p. 1662-1663). This analysis must be rejected if we take outcomes descriptions to include *everything* you care about.

[42] I assume here that there is some small fee that you are prepared to pay to satisfy your preferences—see Cantwell (2003, Section 5) for a discussion of cases where this assumption might fail.

'Up' at any of the second choice nodes since, whatever you do at the first node, you strictly prefer the outcome you get by taking the 'Down' option. But this means that the only plans you can enact are dominated: you either end up with $a - \$1$ when you could have had a, $b - \$1$ when you could have had b, or $c - \$1$ when you could have had c. So, preference cycles violate the Non-Dominated Plan constraint.[43]

Just as with the Allais Preferences, there have been a range of responses to this. Some claim that the agent with cyclic preferences will simply see the loss coming and refuse to do business with the would-be money-pumper (e.g., Schick 1986). But it is unclear what that means in the case just given—if $a - \$1$, $b - \$1$ and $c - \$1$ are each preferred to the status quo, you end up worse off than you need be if you refuse any offers. Others have developed forward-looking choice strategies where each choice at each node is made in light of subsequent choices you might make (see Anand 2009, Section 5 for a summary). But note that *any* strategy that requires you to satisfy your preferences and select 'Down' at the second choice nodes results in your choosing a dominated plan in Cantwell's case.[44]

Just as Machina gave a backwards-looking solution to the dynamic Allais case, Loomes and Sugden (1987) offer a backward-looking strategy to avoid Money Pumps. They note that at the second choice nodes you face a choice between two outcomes (say, a and $c - \$1$), but you have already faced a choice that *could* have yielded a third outcome (b). So even if your preference is $c - \$1 > a$, you need not choose c when offered a choice between a and $c - \$1$ in the above case since (p. 286) 'the set of retrospectively feasible' choices informs what you choose at that time'. At the second choice node, you have already turned down choices that *could* have led to each of $\{a, a - \$1, b, b - \$1, c, c - \$1\}$, so you evaluate your current choice relative to that set.

[43] Note that we might have to tweak that constraint. Each of $\{a, b, c\}$ is itself dispreferred to another in the outcome set, but to convince the defender of cyclic preferences that the Money Pump leaves them worse off than need be, we should not call *these* impermissible outcomes. What distinguishes $a - \$1, b - \$1, c - \$1$ as particularly bad outcomes? Various answers could be given. One plausible explanation is that, say, $a \succ a - \$1$, for any $x \in O$ if $a - \$1 \succ x$ then $a \succ x$, and if $x \succ a$ then $x \succ a - \$1$. We might say that a is *absolutely preferred* to $a - \$1$ (sometimes this relation is called 'strong covering', but I choose absolute preference since it is a natural strengthening of Ahmed's 2017, p. 998 'superpreference'). We might then move from Non-Dominated Plans to:

> **Non-Absolutely-Dominated Plans:** A rational plan is one that is not absolutely dominated by another. (A plan absolutely dominates another if it yields an absolutely preferred outcome in each state.)

A somewhat different analysis of the Money Pump is given by Andreou (2016), who analyses the worseness of $a - \$1$ to a in terms of dimensions of value.

[44] Rabinowicz (2000) and Dougherty (2014) propose Money Pumps that exploit agents who engage sophisticated choice strategies. Ahmed (2017) outlines a strategy that avoids those money pumps. Note however that Ahmed's strategy requires you to choose an absolutely dominated plan in Cantwell's case. I return to this point in Chapter 5.

And as has been widely noted (see, for example, Anand 1993, p. 339), your preference $c - \$1 \succ a$ directly constrains your *pairwise* choices, not what you choose from larger sets. So, you need not choose $c - \$1$ over a *provided* your previous choices inform the set relative to which you evaluate current options. I further discuss, motivate, and defend this approach to choice in Chapter 6.

Some offer still more radical solutions for agents who seemingly face dominated plans. McClennen (1988, 1990) argues that rational agents engage in *resolute choice*. Roughly, when you make a decision you resolve to carry out some plan, thereby constraining your future choices.[45] For example, a resolute choice in the Money Pump might be 'Up, Up'—you thereby constrain your future self to going Up at the second choice node (even if you would otherwise select 'Down' at that node). That rationality requires (or even permits) resolute choice is controversial (though see Buchak 2013, pp. 176-177 for a qualified endorsement).

And still others deny the normative significance of sure losses over time altogether. Levi (2002), for example, argues that dominating sequences are not feasible options. For the agent who makes present choices in light of future ones (more precisely, who engages in *sophisticated choice*), 'picking a and sticking with it when offered c' is simply not a feasible plan (since they know that they will deviate from it), hence not something that they can be rationally criticised for failing to do. (Though, following Steele 2010 p. 474, we might note that such plans are only infeasible *because* of your cyclic preferences, and since the (ir)rationality of such preference is precisely what is up for grabs, we might consider plans feasible if they are feasible relative to preferences you might adopt.) Even more radically, Hedden (2015) defends a view on which there are no genuine diachronic norms. On that view, choosing a dominated plan might be evidence of a synchronic failing, but it does not constitute a normative failings in and of itself.

Cantwell (2003) argues that some, but not all, agents who choose dominated plans are irrational. He argues that dominated plans are bad because they demonstrate that your current preferences are, by your own lights, bad to follow (p. 387). He therefore allows that choosing a dominated plan is permissible if, say, it is the result of changing tastes throughout the decision-making process (as in the famous case of Ulysses and the Sirens). I return to Cantwell's argument in

[45] There are various ways of spelling this idea out. Your future self could have a preference for carrying out resolutions. Or, forming a resolution might change your (future) preferences. Or, you might act contrary to your preferences in order to carry out a resolution. McClennen adopts an approach on which resolution-formation influences, rather than overrides, your preferences. While there are similarities between Machina's, Loomes and Sugden's and McClennen's approaches to dynamic choice—Machina (1989, footnote 26) lists both Loomes and Sugden 1986 and McClennen 1988 as antecedents to his view—there are important normative differences, which I discuss in Chapter 6.

Chapter 5 and conclude that even if you have cyclic preferences that lead you to select dominated plans, this does not show that your preferences are bad by your own lights.

So again, dominance has substantial implications for rationality in the dynamic setting. What constitutes a dominated plan, which preferences result in your choosing dominated plans, and what the normative significance of this is are all up for grabs.

0.3 Conclusion

Just what *is* a dominance principle? Dominance-talk gets used in many and varied ways, with Dominance principles ranging from the (relatively) innocuous to the substantive. I have distinguished between three kinds of principle:

- Dominance principles constraining preferences over acts. This family of principles includes State-wise Dominance, the Sure-Thing Principle, Event-wise Dominance, Decomposability, Betweenness, Mixture Dominance, Stochastic Monotonicity, and First-Order Stochastic Dominance.
- Dominance principles constraining choice. This includes Non-Dominated Choices, Weakly Non-Dominated Choices, and plausibly various Expansion and Contraction consistency requirements.
- Dominance principles constraining plans.

Even this list is incomplete. I have ignored dominance principles constraining preferences over infinite acts—those with infinitely many possible outcomes[46]—and plans involving infinite sequences of acts[47]. And I have failed to even gesture at the myriad ways dominance has been applied outside of decision theory. (To get an idea of how much such a discussion would add, dominance constraints have been used by formal epistemologists support update via Bayesian Conditionalisation (Armendt 1992), Jeffrey Conditionalisation (Armendt 1980), the Reflection Principle (van Fraassen 1984), precise probabilism (Elga 2010), various responses to the Sleeping Beauty Paradox (Briggs 2010a), various norms on the grounds of accuracy arguments (Pettigrew 2015a; Briggs and Pettigrew 2020), and so on.) Nonetheless, I hope it is clear both how much hinges on getting dominance right and how difficult dominance is to pin down.

[46] See, for example, Hájek and Nover (2006), Colyvan (2008) for discussions of such cases that make extensive use of dominance-reasoning.

[47] See, for example, Arntzenius et al. (2004) and Bartha et al. (2014) for discussions of such cases that make extensive use of dominance-reasoning.

It might be a stretch to say that all of decision theory amounts to getting dominance right, but not much of a stretch. Dominance principles are integral in our toolkit for constructing a sensible decision theory. Given the variety of uses such tools have been put to, it would be foolish to try to intervene in every debate. Instead, in what follows I focus on a handful of test cases that reveal the contours (or fault lines, depending on which side of the fence you end up on) of dominance. Some tools will be sharpened, some will have their use restricted, and others will be thrown out. I now open the toolbox.

Chapter 1

Dominance, Decomposability, and Betweenness[48]

1.0 Risk, Betweenness, and Decomposability

Risk matters in a way that renders Savage's Sure Thing Principle (STP) false. We can, however, accept risk-sensitive preferences without giving up on the dominance reasoning that drives Savage's Businessman: if one act does better whatever you might learn, then it does better *simpliciter*. As discussed in the Introduction, we can accommodate the Allais Preferences if we accept Grant et al.'s (2000) Decomposability axiom. And, assuming that you make your decisions relative to a unique probability function over states, Decomposability is equivalent to:

> **Betweenness:** For all acts f, g, if $f \sim g$ then for all probabilities α, $f \sim \alpha f + (1 - \alpha) g \sim g$.

Betweenness will be the focus of this chapter. Decomposability is of course thereby in the background, but since I always work with a unique probability function, any reference to Betweenness can be interpreted as a reference to Betweenness and Decomposability.

Betweenness is a compelling normative constraint. In particular, I argue that interaction effects that justify the Allais Preferences do not justify Betweenness-violations. This means that Buchak's *Risk-Weighted Expected Utility* (REU) gives up too much to accommodate risk-sensitivity. And while Betweenness-satisfying models such as Chew's (1983, 1989) *Weighted Linear Utility Theory* (WLU) are typically interpreted as descriptive models, I demonstrate that WLU can be given a sound normative interpretation and is compatible with everything we might want from a normative theory. Moreover, WLU has independent normative advantages over existing decision theories. For example, it allows for stakes-sensitive risk-attitudes: the way that risk matters to you may depend on the magnitudes of the possible losses and gains you face. By contrast, REU requires that your risk-attitudes be the same whether the worst-case scenario in some gamble is losing your life or winning a yacht. So, Betweenness-satisfying theories like WLU are plausible normative theories that do a better job than existing theories at accommodating pre-theoretic

[48] Sections 1.1-1.5 are drawn from Bottomley and Williamson (Manuscript). Co-authorship was equal, both in terms of initial conception of ideas and writing. Though the material from 1.6 onwards is not drawn from that paper, it has been informed by helpful discussions with my co-author.

intuitions—Betweenness and Decomposability are therefore normatively compelling dominance principles.

1.1 Taking Risks: Orthodoxy

Recall the idea behind Savage's Sure-Thing Principle (STP): if two acts agree in some states, then your preference is settled by what happens on states where they disagree. Given probabilistic sophistication, this corresponds to the *Independence* axiom (see Buchak 2013 p. 171):

> **Independence:** For acts f, g, if $f \succcurlyeq g$ then for all probabilities α and acts h, $\alpha f + (1-\alpha)h \succcurlyeq \alpha g + (1-\alpha)h$.

This says that given a preference between two acts, you retain that preference on 'substituting' or 'mixing' in a third act. The reasoning is that since $\alpha f + (1-\alpha)h$ and $\alpha g + (1-\alpha)h$ both contain equal probabilities of h, we cancel out the contribution of h and compare those acts based solely on f and g. Where STP says that we can ignore events on which acts are identical, Independence says that we can ignore regions of probability space on which acts are identical.

A paradigmatic Independence-satisfying (and STP-satisfying) theory is *Expected Utility Theory* (EUT). If your preferences over outcomes can be represented by a utility function $u: O \to \mathbb{R}$ unique up to positive affine transformation,[49] then define the expected utility of $f = \alpha_1 o_1 + \cdots \alpha_n o_n$ to be:

$$EU(f) = \sum_i \alpha_i \cdot u(o_i)$$

This says that for each outcome that f might yield, weight the utility of that outcome by the probability that f yields that outcome. EUT then says that $f \succcurlyeq g$ just in case $EU(f) \geq EU(g)$. Following Buchak (2013, p.19), one way of thinking about EUT is that an act's instrumental utility is made up of the utility of its outcomes, where the 'contributive value' of each outcome is its probability-weighted utility.

EUT raises several questions: when is a unique u guaranteed to exist, and what kinds of agents can be represented as EU-maximisers? Both questions are answered by von Neumann and

[49] A utility function represents your preferences if $u(x) \geq u(y)$ just in case $x \succcurlyeq y$. Such a function is unique up to positive affine transformation if for any u' that also represents preferences, there exist $a, b \in \mathbb{R}$, $a > 0$ such that $u'(\cdot) = a \cdot u(\cdot) + b$.

Morgenstern's celebrated *Representation Theorem*, which shows that if an agent satisfies the following four axioms, then they can be represented as maximising EU:[50]

Completeness: For all f, g, $f \succcurlyeq g$ or $g \succcurlyeq f$.

Transitivity: If $f \succ g \succ h$, then $f \succ h$; if $f \succcurlyeq g \succcurlyeq h$, then $f \succcurlyeq h$.

Continuity: If $f \succcurlyeq g \succcurlyeq h$, then there exists an α such that $g \sim \alpha f + (1 - \alpha)h$.

Independence: If $f \succcurlyeq g$ then for all α and acts h, $\alpha f + (1 - \alpha)h \succcurlyeq \alpha g + (1 - \alpha)h$.

So, according to EUT, rationality consists in obeying the vNM axioms, which is equivalent to maximising EU. Note that we might take Completeness, Transitivity, and Continuity as working structural assumptions—if so, then we can think of Independence is *the* defining structural feature of EUT.

1.2 Taking Risks Seriously: Alternatives to Orthodoxy

EUT is an elegant theory, but it places a high bar on rational behaviour. Because it satisfies Independence, it rules out the Allais Preferences. Recall from the Introduction:

	$1 - 89$	90	$91 - 100$
a	\$1 million	\$1 million	\$1 million
b	\$1 million	\$0 million	\$5 million

	$1 - 89$	90	$91 - 100$
c	\$0 million	\$1 million	\$1 million
d	\$0 million	\$0 million	\$5 million

Many here prefer $a \succ b$ and $d \succ c$, though Independence and Savage's STP rule this out—your preferences should not depend on 'common consequences', or what happens on Tickets 1-89.

A less widely discussed counterexample to Independence (at least among philosophers), though perhaps a simpler and more compelling one, is:

[50] The following axioms are actually due to Marschak (1950). Von Neumann and Morgenstern's initial presentation is equivalent but less intuitive.

> **Common-Ratio:** Jane likes London but loves Rome. She strictly prefers going to London for sure over taking a risk on an 80% chance of going to Rome (and a 20% chance of going nowhere). On the other hand, she strictly prefers a 20% chance of going to Rome (and an 80% chance of going nowhere) over a 25% chance of going to London (and a 75% chance of going nowhere).

Jane's preferences are sensible enough and can be rationalised in much the same way as the Allais Preferences. Given a high probability of going to London, the increased chance of losing out on a holiday altogether looms large. So, Jane plays it safe when the background conditions are favourable. On the other hand, when the probability of getting a holiday is low, she is happy to bear more risk and take slightly worse odds on a Roman holiday.[51]

While reasonable, Jane's risk-sensitive preferences violate Independence. To see this, let lotteries between Rome, London, and No Holiday be represented by triples of the form $(p_{\text{Rome}}, p_{\text{London}}, p_{\text{None}})$. Jane's preferences are then:

1. $(1,0,0) \succ (0,1,0)$
2. $(0,1,0) \succ (0.8, 0, 0.2)$
3. $(0.2,0,0.8) \succ (0,0.25,0.75)$

Independence requires that Jane's preferences remain fixed when we substitute in a third lottery. So, 2. remains fixed when we substitute in a 75% chance of nothing:

$$.25(0,1,0) + .75(0,0,1) \succ .25(0.8,0,0.2) + .75(0,0,1)$$

Which is equivalent to:

$$(0,0.25,0.75) \succ (0.2,0,0.8)$$

And this contradicts 3. So, Jane's reasonable preference violate Independence, and hence EUT.

From here on, I assume that the Common Ratio and Allais Preferences are rational. We have already seen in the Introduction that one powerful argument against such preferences—

[51] Some might worry that the shift from *certainty* to lack of certainty is doing a lot of work here, and we should write off Jane's preferences because of the certainty effect. If so, note that we might replace Jane's preferring London to an 80% shot at Rome with her preferring a 95% chance of London to a 76% shot at Rome—Jane's preferences are still reasonable and violate Independence, though no shift away from certainty is involved. We must also be careful when invoking the 'certainty effect'. Strictly speaking, effects are descriptions of behaviour, not explanations of behaviour. When saying that the certainty effect explains some patterns of preferences, this raises both the question of what psychological mechanisms underwrite that effect, and the question of whether our best theory of rationality allows for that effect. However we label Jane's preferences, we can ask whether those preferences are compatible with a normatively plausible decision theory.

coherence between your conditional and unconditional views—fails. The Allais Preferences do not invalidate the plausible reasoning that drives Savage's Businessman. Several other arguments against risk-sensitive preferences have been given.[52] Here, I simply state that I do not find such arguments persuasive. And there have been several attempts to show that EUT can accommodate canonical risk-sensitive preferences.[53] And again, though it deserves a fuller treatment, I here simply side with those who think that EUT is incompatible with risk-sensitivity.

If we must reject EUT, then we need an alternative framework. Lara Buchak (2013, Chapters 2 and 3; 2014) has worked out the details of an alternative normative model, *Risk-Weighted Expected Utility Theory*. Buchak adopts a rank-dependent approach—rational agents might agree about the utility and probability of outcomes but disagree about the importance of, say, securing a higher minimum compared to a higher maximum when comparing lotteries. Such risk-attitudes are represented by a function $r: [0,1] \to [0,1]$ such that $r(0) = 0, r(1) = 1$, and r is non-decreasing on $[0,1]$.[54] REU requires that an act's possible outcomes are ordered so that $o_1 \preccurlyeq o_2 \preccurlyeq \cdots \preccurlyeq o_n$. Let $S_{f \geq o}$ denote the set of states such that f yields an outcome weakly preferred to o on those states. Then the *risk-weighted* expected utility of f is defined:

$$REU(f) = u(o_1) + \sum_{i=2}^{n} r(C(S_{f \geq o_i})) \cdot [u(o_i) - u(o_{i-1})]$$

And Risk-Weighted Expected Utility Theory (REU) says that $f \succcurlyeq g$ just in case $REU(f) \geq REU(g)$. Think of REU as telling you to weight how much of an improvement o_i is (over the next worse outcome) by the risk-weighted probability of doing at least as well as o_i. Clearly, each act is ranked at least as well as its worst outcome. And if r is convex, then the REU of a lottery will always be lower than its EU (cf. Buchak 2013, p. 1099). So, REU with a convex r satisfies a simple definition of risk-aversion: each lottery is valued below its mean—you always prefer a sure thing of some outcome over a risky act with equivalent expected payoff.

REU accommodates the Allais Preference as follows. Say that $u(\$x) = \sqrt{x}$ and take the convex $r(p) = p^2$. Then:

[52] See *inter alia* Hammond (1977, 1988), Briggs (2015), Pettigrew (2015b), Ahmed (2016), and Thoma (2019).

[53] See *inter alia* Broome (1991 Chapter 5), Bradley (2017, Section 9.5.1), Stefánsson and Bradley (2019), and Weirich (2020). The most prominent strategy involves *redescription* of outcomes. For example, Broome treats 'getting nothing' and 'getting nothing when you were likely to get a million' as distinct outcomes in Allais. I agree with Buchak (2013 p. 123) that an agent can have the Allais Preferences without the re-individuated outcomes 'being legitimate descriptions of the choice problems he takes himself to face'.

[54] A non-decreasing r may violate First-Order Stochastic Dominance but will satisfy Weak First-Order Stochastic Dominance. We could insist on an increasing r, which guarantees First-Order Stochastic Dominance.

$$REU(a) = u(\$1m) = 1{,}000$$

$$REU(b) = u(\$0) + .99^2[u(\$1m) - u(\$0)] + .1^2[u(\$5m) - u(\$1m)] \approx 992.5$$

$$REU(c) = u(\$0) + .11^2[u(\$1m) - u(\$0)] \approx 12.1$$

$$REU(d) = u(\$0) + .1^2[u(\$5m) - u(\$0)] \approx 22.4$$

This gives $a \succ b$ and $d \succ c$ as required. Because of the way r transforms probabilities of ranked outcomes, the introduction of a small probability of $\$0$ is not offset by the introduction of $\$5$ million in the first pair of lotteries, while it is offset in the second. In recent years, this rank-dependent approach has received a lot of attention and perhaps deserves to be called the orthodox alternative to orthodoxy for dealing with risk.[55]

1.3 Risk and Irrationality

REU accommodates risk-sensitive preferences but violates Betweenness. To illustrate, consider the outcomes $\{\$1, \$4, \$25\}$, again with $u(\$x) = \sqrt{x}$ and $r(p) = p^2$. Each of the following lotteries has REU of 2:

Gamble A: $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$

Gamble B: $(0,1,0)$

That is, our REU-maximiser judges a fair coin-toss on nothing and $\$25$ to be precisely as good as $\$4$ for sure. Now consider:

Gamble C: A fair coin toss to decide between Gamble A and Gamble B.

This has REU:

$$REU(C) = REU\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) = 1 + \frac{3}{4} < 2$$

[55] Interestingly, if we adopt a power-weighting function, $r(p) = p^n$, then REU is incompatible with the Common Ratio preferences. This means that while the canonical $r(p) = p^2$ is extremely risk-averse (indeed, it is common to define degree of risk-aversion in terms of convexity of r), it only accommodates a limited kind of risk-aversion. Diecidue et al. (2009) show that all and only risk-weighting functions of the form $r(p) = p^n$ avoid Common Ratio preferences.

So, REU violates Betweenness, even for very simple lotteries. This case demonstrates how odd Betweenness-violations are. Firstly, tossing a coin to decide between indifferent options strikes me as a common-sense thing to do, as does closing your eyes and picking at random. But an agent who tosses a coin or picks randomly effectively chooses Gamble C over A or B. Moreover, note that Gamble C is a coin-toss between Gamble A and *its own certainty equivalent*. So, REU tells you that A's risks and benefits are precisely offset by a sure thing of $\$4$, but then tells you that some probability of A or that very certainty equivalent is *worse* than A. Odd indeed! $r(p) = p^2$ is a paradigmatically risk-averse weighting function, yet it says that introducing some probability of the certainty equivalent makes things worse. Intuitively, introducing some chance of the certainty equivalent, if anything, makes things less risky, not more. Indeed, Gamble A has a 50% chance of yielding nothing, while Gamble C only has a 25% chance of yielding nothing—Gamble C has moved probability away from extreme outcomes towards the certainty equivalent. And yet REU says that the paradigmatically risk-averse agent will judge this to be a bad thing. This should all make us question whether REU really captures rational risk-aversion.

1.4 **Defending Betweenness**

The mere fact that REU violates Betweenness is not a decisive reason to reject REU. Firstly, defenders of REU might be able to tell some story that rationalises Betweenness-violations and so convince us that Betweenness is overly stringent (despite its *prima facie* plausibility). Secondly, defenders of REU might be able to point to advantages of REU and argue that, even if Betweenness-violations are odd, REU is the best way of accounting for risk-aversion, all things considered. In this section and the next, I show that neither of these are the case.

Begin with the first point, that we might be able to rationalise Betweenness-violations. There has been no systematic defence of Betweenness from a normative perspective. Indeed, Camerer and Ho (1994, Section 1.2) claim that the intuitive force of Betweenness is really just the intuitive force of Independence—it is a kind of substitution invariance constraint, and Allais and Common Ratio Preferences show that we have good reason to reject such constraints.[56]

[56] Grant et al. (2000) consider Decomposability and Betweenness to be normatively compelling. Their argument is that *because* those principles vindicate Savage's Businessman, they are normatively compelling. This seems to be best construed as an argument from reflective equilibrium: we have some piece of plausible reasoning (Savage's Businessman), and if we can systematically accommodate deeply engrained intuitions (such as the Allais Preferences) *without* rejecting that reasoning, then we should. This does not, however, deal with the point from Camerer and Ho—that once we reject Independence and STP, the intuition driving Savage's Businessman disappears. In what

But I think we can give an argument for Betweenness that does not appeal to the kind of reasoning that underwrites Independence. Firstly, note how natural Betweenness-reasoning is. In the above case, the reason that Gamble C is worse than B must be because of the possibility of \$1. But C introduces \$1 *and an equal chance of* \$25. And the initial preference $\text{Gamble A} \sim \text{Gamble B}$ tells us that the agent is prepared to sacrifice \$4 for equal chances of \$1 and \$25. So, it is incoherent to rank Gamble C as below A when the agent has already said they are prepared to introduce some probability of \$1 for equal chances of \$25. This is a good first-pass motivation for endorsing Betweenness: in declaring a certainty equivalent for some gamble, you thereby declare what chances of loss you are prepared to live with for some chance of gain. And that thought does not presuppose Independence.

Very well, the defender of REU might reply. Risk-sensitive preferences tell us that there can be *interaction effects* between outcomes. Just because you judge \$4-for-sure to be as good as a coin toss on \$1 and \$25, that does not mean that you must judge \$4 to be offset by equal chances of \$1 and \$25 in *all* lotteries. So, you are not paying for nothing if you pay to avoid randomisation—you are getting a lottery with different global properties. (The phrase 'global properties' refers to any structural feature of a gamble not reducible to its mean—this includes minimum, maximum, variance, and so on.)

To defend Betweenness in light of this concern, consider the following case (like all good arguments, it involves opaque and transparent boxes):

> *Three Boxes*: You get to choose one of the three boxes in front of you. Two are transparent and one is opaque. Transparent Box A contains lottery Ticket A. Transparent Box B contains lottery Ticket B. You are indifferent between Ticket A and Ticket B. Inside the Opaque Box is another copy of either Ticket A or Ticket B, determined by a fair coin-toss.[57]

	Coin Heads	Coin Tails
Box A	A	A
Box B	B	B
Opaque Box	A	B

follows, I show (contra Camerer and Ho) how Betweenness can be given normative justification without appealing to Independence, which I take to complement Grant et al.'s reflective equilibrium methodology.

[57] I assume the coin is fair, but nothing substantive hinges on this.

Anyone who violates Betweenness here by (dis)preferring the Opaque Box owes us some story. They can have A or B outright, so why (dis)prefer to randomise?

Recall that we can violate Independence because of interaction effects: your preference between f and g is not decisive in determining your preference between $\alpha f + (1-\alpha)h$ and $\alpha g + (1-\alpha)h$. This is because a $(1-\alpha)$ probability of h affects how αf and αg contribute to your overall evaluation of gambles. The fact that you *could* receive h with probability $(1-\alpha)$ affects how you assess the risks and benefits of an α probability of receiving f or g. Independence incorrectly requires that what happens on the $(1-\alpha)$-region of probability space have no bearing on how you evaluate risks borne on the α-region of probability space. Some such reasoning may explain why it is rational to violate Betweenness in Three Boxes—perhaps the fact that you could have either A or B affects the contributive value of the outcomes of the other.

But such a story is mysterious. Why? Say that you take the opaque box in Three Boxes. Then you are guaranteed to get either A or B, both of which you are prepared to choose *when you could have the other*. If the opaque box contains A, you get something you were perfectly happy to take outright (A when you could have had B). If the opaque box contains B, you also get something you were perfectly happy to take outright (B when you could have had A). There can therefore be no 'rational dissatisfaction' on discovering what the box contains—you are simply getting something you were happy to have while not getting something you were prepared to turn down. Indifference between A and B means that you do not care which way the uncertainty is resolved.

Contrast this with Allais, a paradigmatic case where interaction effects *can* affect your overall preferences. In that case, finding whether one of Tickets 1-89 or Tickets 90-100 is drawn makes the world of difference. In the first pair of Allais gambles, learning that one of Tickets 1-89 was drawn is great news (you are guaranteed to be a millionaire), while in the second pair learning that one of Tickets 1-89 was drawn is terrible news (you are guaranteed to be a pauper). So, what happens on the 1-89 region of probability space affects how you evaluate what happens on the rest of probability space. If 1-89 is bad, you might take more of a risk on the remaining tickets. If it is good, you may take less of a risk on those tickets. The crucial feature is that what happens in Tickets 1-89 is relatively good or bad.

Three Boxes lacks this feature. Finding out how the coin landed carries no good or bad news. Because you are indifferent between the resolutions of event-wise uncertainty, you have no reason to evaluate one region of probability space differently. You are simply finding out which of two options you receive, and you are happy to trade between those options in the risk-free context. So, Betweenness-violations cannot be naturally explained in the same way as the Allais Preferences.

Let's say the defender of REU digs their heels in and insists that the opaque box really is worse than either transparent box. They might be reasoning as follows: 'If I get Ticket A, the fact that I *could* have B with probability $.5$ makes a difference to how I evaluate $.5A$'. But that is close to incoherent! They treat A differently when they could have B, though they are perfectly happy to take A when they can have B. Similarly, they are prepared to take B when they can have had A. The fact that our Betweenness-violator is prepared to trade between A and B means that interaction effects which purport to justify Betweenness-violations are mysterious. Betweenness-violations cannot be given the same intuitive justification as Independence-violations can, and it is implausible that agents are prepared to trade between two lotteries but not randomise over them.[58]

Of course, the defender of REU can say that there *might* be interaction effects that explain Betweenness-violations, even if it hard to say precisely what they are or why they matter. REU is a well-axiomatized theory, so we could simply define normatively significant interaction effects as the interaction effects that REU vindicates. But at this stage, we are owed a story about precisely what motivates this claim. Accepting that global properties matter does not commit us to thinking that *all* global properties matter. And it does no explanatory work to say that the global properties that matter are the ones that REU picks out.[59] *Three Boxes* underwrites Betweenness and shows that Betweenness-violations are importantly different to Independence-violations.

1.5 <u>Risk Without Irrationality</u>

I have argued that Betweenness-violations are mysterious, so we have no principled story to tell that rationalises Betweenness-violations. I now turn to the second move the defender of REU

[58] Note Corollary 6 in Chew (1983) entails that we can construct a Three Box case in which the opaque box has the same expectation as the transparent boxes. So, things are even worse for the defender of REU: they have to explain why you should (dis)prefer the Opaque Box in cases where it has the very same mean as the transparent boxes.

[59] Buchak can of course respond that by randomising over lotteries, the rank of each outcome changes. But if this is to explain Betweenness-violations, we need to say why we should care about rank in the first place. I personally have no strong intuition that rank considerations explain my intuitions in the Allais case—and we will see in Sections 5 and 6 that rank-dependent theories have strange normative upshots, which makes me suspicious that a rank-dependent view is strongly pre-theoretically motivated.

could make—they could claim that while it is hard to give a principled defence of Betweenness-violations, such violations are nonetheless all-things-considered permissible because REU has advantages that rival theories do not.

I argue that this is not the case. I can think of two advantages that REU might have over rival theories:

i) Rival theories violate some constraint at least as plausible as Betweenness.
ii) Rival theories cannot be given a good normative interpretation.

I argue that an appropriately modified version of Chew's *Weighted Linear Utility Theory* (WLU) addresses both worries. While extant versions of WLU permit violations of plausible normative constraints (such as Weak First-Order Stochastic Dominance), a strengthened version of WLU respects all widely agreed upon normative criteria. I then show, drawing on Buchak's own interpretation of REU, that WLU can be given a sound normative interpretation.

1.5.1 Risk and the Non-Negotiable Features of Rationality

Not a lot is widely agreed on in decision theory. But, following Buchak (p. 37), you might think that any theory which deserves to be called normative must at least:

- Respect Weak First-Order Stochastic Dominance (FOSD).
- Be compatible with Second-Order Stochastic Dominance (SOSD).
- Satisfy Transitivity.

I take these desiderata for granted in this chapter. Weak First-Order Stochastic Dominance, as we have seen, is a basic criterion for one lottery being better than another. Second-Order Stochastic Dominance can be spelled out in many ways, but most helpful here is that a second-order stochastically dominates b if b is a *mean-preserving spread* of a. For finite lotteries this is equivalent to the condition that b can be reached from a via a sequence of transformations that shift pairs of probability mass away from the mean while leaving the mean unchanged. (See Proof of Theorem 2 in the appendix of this chapter for a formal statement).[60] I say that you

[60] Rothschild and Stiglitz (1970) provide an early statement of SOSD and alternative characterisations of that principle. Cohen (1995) discusses various characterisations of risk-aversion. Typically, we talk about SOSD with respect to some good (for example, you satisfy SOSD with respect to money if you prefer not to shift probabilities of monetary outcomes away from the mean while leaving the expected monetary payoff the same). Since we are dealing with pure risk-aversion—risk-aversion with respect to utility and not some good like money—I always take SOSD to be with respect to utilities. That is, a second-order stochastically dominates b if b can be reached from a

satisfy SOSD just in case you prefer a to b whenever a second-order stochastically dominates b. Compatibility with SOSD is more controversial than FOSD, but it is a standard definition of risk-sensitivity, so I take it for granted here. If you satisfy SOSD, you are averse to shifting probability of outcomes away from the mean, all else being equal. Transitivity is a widely accepted structural principle. Together, I call these three conditions the 'normative core'.[61]

Chew's (1983) Weighted-Linear Utility Theory is a paradigmatic Betweenness-satisfying theory that is compatible with the normative core. That such a theory exists shows that we do not face the kind of trade-off among constraints that would require us to abandon Betweenness.

Like REU, WLU introduces a weighting function in addition to utility and probability. Unlike REU, WLU's weighting function is on *outcomes*, $w: O \rightarrow (0, \infty)$. Crucial is that w's range is positive, and that indifferent outcomes are assigned equal weight. This weighting function will allow us to rationalise risk-sensitivity without transforming probabilities as does REU.[62]

For act $f = \alpha_1 o_1 + \cdots + \alpha_n o_n$, define its *weighted-linear utility* as:

$$WLU(f) = \sum_i \frac{w(o_i)}{\sum_j w(o_j)\alpha_j(x_j)} \cdot \alpha_i(x_i) \cdot u(x_i)$$

WLU says that $f \succcurlyeq g$ if and only if $WLU(f) \geq WLU(g)$ and says that permissible options are those that maximise WLU.

The term $\frac{w(o_i)}{\sum_j w(o_j)\alpha_j(x_j)}$ I call o_i's *relative weight* in f—the outcome's weight divided by the average weight of the lottery. Think of WLU as weighting each outcome's utility by its probability *and* its relative weight. (More on precisely what this means in the next section.) An outcome whose relative weight is greater than 1 (i.e., whose weight is greater than average) has greater contributive value than its probability-weighted utility, while an outcome whose relative weight is less than 1 has less contributive value than its probability-weighted utility. Note that when w is constant, WLU reduces to EUT, so EUT is a special case of WLU (just as EUT is a special case of REU).

via a sequence of transformations that shift pairs of probability mass away from a's *expected utility*, while leaving that *expected utility* unchanged.

[61] In later chapters, I drop Transitivity. Nonetheless, I grant Transitivity here for the sake of argument. So, even *if* we insist on Transitivity, Betweenness-satisfying theories retain their advantage over Betweenness-violating ones.

[62] My presentation of WLU differs from Chew(1983) here, partly because I provide an explicit functional form for w, which Chew does not, and partly because the acts I discuss are finite, while Chew deals with acts that induce continuous probability distributions.

Chew (1983, Corollary 6) shows that there exist weight and utility functions that are compatible with First- and Second-Order Stochastic Dominance. One wrinkle: not all combinations of weight and utility functions satisfy Weak First-Order Stochastic Dominance. No matter—we can insist that rationality requires WLU maximisation only relative to a pair of u, w that satisfy FOSD (the u,w I give below respect both FOSD and SOSD).[63] We might call this strengthened decision theory WLU*, but I will simply refer to it as WLU in what follows and take FOSD as implicit.

From an axiomatic perspective, WLU weakens just the Independence axiom (see Chew 1989).[64] We retain Completeness, Transitivity, and Continuity, and replace Independence by:

> **Weak Independence:** For all acts f, g, if $f \sim g$ then for every probability $\alpha \in [0,1]$ there is exists a probability $\beta \in [0,1]$ such that for every h: $\alpha f + (1 - \alpha)h \sim \beta g + (1 - \beta)h$.

One way of thinking about Weak Independence is that for any two indifferent lotteries, α-substitutions of the first are always indifferent to *some* fixed β-substitution of the second. Independence is the special case where for each α, $\alpha = \beta$. In the presence of the other axioms, Weak Independence trivially entails Betweenness. I will not defend Weak Independence here, in part because it is one of many possible axioms underwriting both WLU and many other Betwenness-satisfying theories. What matters is that WLU is well-axiomatised, respects the normative core, and as we will see has some attractive properties. A normative comparison between WLU and other Betweenness-satisfying theories is for future work.

To illustrate WLU in action, consider again $u(\$x) = \sqrt{x}$ and a weight function that is everywhere decreasing on $[0, \infty)$:

$$w(\$x) = \frac{1}{1 + \sqrt[4]{x}}$$

Intuitively, an agent with this weight function places greater significance on worse outcomes compared to better ones—rags loom large in comparison to riches. Now consider lotteries over the following outcomes:

[63] Alternatively, we could insist on Stochastic Monotonicity as an axiom in addition to the axioms that Chew uses to underwrite WLU. Again, while Chew for descriptive purposes will allow violations of this axiom, we can supplement the WLU axioms for normative purposes.

[64] See Fishburn 1988 (p. 64) for a discussion of various axiomatisations of WLU.

$$\{\$0, \$100, \$10\,000, \$20\,000\}$$

Note the contrast between WLU and EUT when considering coin-tosses on the *worst* two outcomes:

$$WLU\big((.5, .5,0,0)\big) \approx 1.937$$

$$EU\big((.5, .5,0,0)\big) = 5$$

Recalling that $u(\$x) = \sqrt{x}$, the EU-maximiser pays up to $\$25$ for this lottery. The WLU-maximiser, however, is significantly more cautious: they pay only up to $\$3.75$.

On the other hand, contrast WLU and EUT in the case of a coin-toss on the *best* two outcomes:

$$WLU\big((0,0, .5, .5)\big) \approx 119.071$$

$$EU\big((0,0, .5, .5)\big) \approx 120.711$$

The WLU-maximiser pays up to $\$14{,}177.80$ for this lottery, while the EU-maximiser pays only *slightly* more, $\$14{,}571.07$. This illustrates an interesting feature of the above weight function: because w approximates a constant function for large x, our WLU-maximiser 'looks like' an EU-maximiser when the worst-case scenario is good and the outcomes are not too spread out. Things change, however, on incorporating some probability of a bad outcome:

$$WLU\big((.1,0, .45, .45)\big) \approx 51.345$$

$$EU\big((.1,0, .45, .45)\big) \approx 108.64$$

In response to including a 10% probability of nothing, the WLU-maximiser dramatically reduces their valuation to $\$2{,}636.35$, while the EU-maximiser is far more tolerant to risk and pays up to $\$11{,}802.56$. While this weight function is merely illustrative, it has some appealing features: a high degree of responsiveness to bad outcomes coupled with an almost risk-neutral attitude towards safe gambles. This is because our w sharply increases as we approach $\$0$ from the right, so relative weight (and hence WLU) is more sensitive to increases in probability of outcomes close to $\$0$ compared to outcomes further away from $\$0$.

Our chosen u,w rationalises the Allais Preferences. Using slightly different notation, we can represent that case as:

	$u(x)$	$w(x)$	$p_a(x)$	$p_b(x)$	$p_c(x)$	$p_d(x)$

$\$0$	0	1	0	$.01$	$.89$	$.9$
$\$1m$	1000	$.031$	1	$.89$	$.11$	0
$\$5m$	2236.068	$.021$	0	$.1$	0	$.1$

And we calculate:

$$WLU(a) = 1000$$

$$WLU(b) \approx 810.935$$

$$WLU(c) \approx 3.774$$

$$WLU(d) \approx 5.133$$

This gives $a \succ b$ and $d \succ c$ as required.

Say that function $z: O \to \mathbb{R}$ is *increasing* if $z(o_1) > z(o_2)$ if and only if $o_1 > o_2$, and that it is *decreasing* if $z(o_1) < z(o_2)$ if and only if $o_1 > o_2$. We can also show that our u, w satisfies First- and Second-Order Stochastic Dominance by establishing the following two results:[65]

> **Theorem 1:** If u and $u \cdot w$ are increasing and w is decreasing, then WLU satisfies First-Order Stochastic Dominance.
>
> **Theorem 2:** If u is increasing, w is decreasing, and $k(x, y) = \frac{u(x)-u(y)}{w(x)-w(y)}$ is decreasing in both x and y, then WLU satisfies Second-Order Stochastic Dominance.

See the Appendix at the end of this chapter for proofs. The intuition behind both Theorems is that in WLU weighted-utility, the function $u \cdot w$, plays a similar role to u in EUT. Just as in EUT the defining property of FOSD is u's being increasing, in WLU the defining property of FOSD is $u \cdot w$'s being increasing. And Theorem 2 introduces $k(x, y)$, which is a measure of how much changes in u are outweighed by changes in w; this is related to the concavity of $u \cdot w$, and note that concavity of u (with respect to, say, money) is the defining property of SOSD (with respect to, say, money) in EUT.

[65] Note that Chew (1983) also provides sufficient conditions for First- and Second-Order Stochastic Dominance. His proofs are for continuous, non-finite acts, whereas I am working in the discrete case. His proofs rely on fairly heavy-duty mathematical tools. The proofs here rely on standard (though tedious!) algebra, so hopefully give a general audience the idea of the structure of WLU.

For $u(\$x) = \sqrt{x}$ and $w(\$x) = \frac{1}{1+\sqrt[4]{x}}$, it is easy to check (by computing first derivatives of $f(x) = \frac{u(x)}{w(x)}$ and partial derivatives of $g(x, y) = \frac{u(x)-u(y)}{w(x)-w(y)}$) that the conditions in Theorems 1 and 2 are met, so our canonical u,w satisfies FOSD and SOSD. And since Transitivity is one of the WLU axioms, our chosen u, w respects the normative core.

We have a plausible Betweenness-satisfying theory, WLU, that satisfies everything that Buchak insists on from a decision theory. Note in particular that we can respect Second-Order Stochastic Dominance, a very general characterisation of risk-aversion, without violating Betweenness. This strengthens my previous claim that we cannot tell an intuitive story to rationalise Betweenness-violations, since whatever interaction effects or global properties we invoke to explain Betweenness-violations, they are not related to *risk* in a systematic way. Violations of Betweenness do not characterise risk-sensitivity but some other kind of sensitivity, call it randomisation-sensitivity.

1.5.2 What are Weights?

Weight functions allow us to rationalise risk-sensitivity. But what *is* w? As stated above, we might still adopt a Betweenness-violating theory if rival theories cannot be given a sound normative interpretation. For WLU to serve as a candidate normative theory, w had better be more than a formal fix.

I argue that we can interpret w as a normatively relevant *significance* function. I get at this idea by first considering two natural interpretations of w that would disqualify WLU as a normative theory.

Firstly, you might interpret relative weights as measuring the degree to which an outcome has some intrinsically (dis)valuable property. Outcomes with relative weight greater than 1 might induce (intrinsically valuable) elation due to thrill, surprise, and so on. Similarly, outcomes with relative weight less than 1 might induce (intrinsically disvaluable) disappointment due to regret, remorse, anxiety, and so on. Widely accepted is the idea that the utility function is supposed to capture everything you intrinsically care about.[66] Broome (1991, p. 103) argues that two outcomes are distinct if and only if they differ with respect to properties that you rationally care

[66] Note also Buchak's argument that the Allais Preferences should not be rationalised by intrinsic concern for global properties (Section 1.4.1), or by the utility of outcomes being sensitive to global properties.

about; Pettit (1991) spells this out in terms of those outcomes differing with respect to some desirable property. On such views, relative weights cannot be a measure of some normatively significant property: the one outcome can have *different* relative weights in different lotteries, meaning that one outcome counts as two outcomes—which is incoherent.[67]

A second interpretation is due to Camerer (1989, p. 66), who interprets relative weights as subjective distortions of probabilities. The WLU-maximiser is therefore guilty of misrepresenting probabilities by treating outcomes with relative weight greater than 1 as more likely than they actually are (and outcomes with relative weights less than 1 as less likely than they actually are). It may be that global properties do interact with biases such that we sometimes misrepresent probabilities. But if such biases are the only explanation for risk-sensitive preferences, then the rational thing to do is revise those preferences. (We could imagine a similar argument related to utilities: relative weights measure your disposition to misrepresent values, reasons, and so on. Again, the rational response to such misrepresentation would be to revise your preferences.)

There is an important analogy with the development of REU here. Prior to Buchak (see 2013, pp. 43-7), many interpreted the third component of rank-dependent theories (Buchak's r) as a measure of optimism and pessimism.[68] One of Buchak's important innovations is to interpret r as weighting undistorted probabilities such that each outcome's contributive value depends on something other than its probability and utility. This is right: normative departures from EUT should *not* involve psychological distortions of probabilities, but a third component distinct from utility and probability.

We can therefore think of w as representing the *significance* of outcomes, or a measure of how much you are concerned with securing some outcomes compared to others. Rational agents can agree on the desirability and probability of outcomes but still disagree about the instrumental utility of acts because they disagree about the relative significance of the outcomes involved. Buchak expresses a similar point (italics mine):

> '… it is plausible to think that some people are more concerned with the worst-case scenario than others, again, for purely instrumental reasons: because they think that *guaranteeing* themselves something of moderate value is a better way to satisfy their general aim of getting some of the things that they value than is making something of

[67] If we deny that outcome descriptions are exhaustive, we could take u to measure the contribution of an outcome's *intrinsic utility* (the component of your outcome-evaluation fixed independently of what else you could have had) that is then transformed by relative weight, which represents something like *comparative utility* (the component of your outcome-evaluation settled what else you could have had). Though I do not pursue this line here, thanks to Eleanor Boxall for this suggestion.

[68] Quiggin (1982, p. 333) takes r to represent an agent's attitudes *to* probabilities, rather than as capturing the way that probabilities themselves determine contributive value.

> very high value merely possible … Alternatively an agent might be concerned with the best-case scenario: the maximum weighs more heavily in the estimation of a gamble's value than the minimum does, even if those two outcomes are equally likely … Thus, in addition to having different attitudes towards outcomes and different evaluations of likelihoods, *two agents might have different attitudes towards some way of potentially obtaining some of these outcomes*.' (Buchak 2013, p. 49)

On Buchak's view, rational agents can have different attitudes towards various ways of realising the same goals. That is correct: instrumental rationality is permissive enough to allow an outcome with fixed probability and utility to have different contributive values in different lotteries. But there is an important difference between the view I am defending and Buchak's. For Buchak, an agent's risk-attitude concerns how important it is to avoid the *worst* ranked outcome, seek the *best* ranked outcome, and so on. Words like 'worst' and 'best' here are purely ordinal: they refer to the ranking of outcomes in the specific lottery under consideration. The WLU-maximiser, however, does not care about avoiding the worst outcome simply because it is the *worst* outcome in the lottery, seeking the best outcome simply because it is the *best* outcome, and so on. Rather, because w takes outcomes as its domain, the relative weight of an outcome in a lottery depends on the *specific* outcomes that lottery might yield. The WLU-maximiser cares more or less about securing certain *outcomes* relative to others, so the contributive value of an outcome in a lottery will depend on richer information than rank-order. My illustrative w above is everywhere decreasing, so it will always attach greater significance to worse outcomes. But the relative weight of the worst outcome in a lottery will depend on the significance of the other possible outcomes, which varies with the specific outcomes in the lottery.

My disagreement with Buchak therefore boils down to a disagreement about *how* rational agents structure 'the potential realisation of some of [their] aims'. Buchak takes this structuring of the possible realisation of ends to depend on the rank-order of those ends, *whatever those ends are*. It does not matter whether the worst outcome is losing your life or winning a yacht. I, however, take the structuring of the potential realisation of your ends to depend on the *specific* ends involved in the lottery, not just their rank. In the next section, I show that this overcomes a serious normative shortcoming of rank-dependent theories.

So, we can interpret WLU as a plausible normative theory. Of course, I have not shown that WLU is the only Betweenness-satisfying theory that can be given such an interpretation (Gul's 1991 *Disappointment-Aversion* Theory is another promising model to explore). But crucial for present purposes is that we have a Betweenness-satisfying theory, that it satisfies our normative core, and that it can be given a plausible normative interpretation. We should therefore have no qualms about trading in a Betweenness-violating model for a Betweenness-satisfying one.

1.6 Stakes Sensitivity

It often seems appropriate to take a risk on some low-stakes gamble while refusing to take a risk on a scaled-up version of that same gamble. Many of us would accept a coin-toss that yields a small win or loss without accepting a coin-toss that yields the same win or loss a thousand times over. More generally, you might be risk-averse in high-stakes situations (e.g., when buying insurance) while being more risk-tolerant in low-stakes situations (e.g., when buying vegetables). Existing theories have been criticised for not allowing this (see Armendt 2014 and Hájek Forthcoming[a]). For example, Armendt discusses fair coin-tosses involving different stakes (one on hundreds of dollars, the other on billions) and notes the strong intuition that our risk-attitudes in low-stakes scenarios need not settle our preferences in high-stakes scenarios. The issue is that REU fixes a single risk-attitude (a single attitude towards, say, the importance of securing a higher minimum) that applies to any lottery you might face.

Armendt then notes:

> '[EUT] is what we get from REU when r is the identity function $r(p) = p$ … So it may seem inappropriate to ask, why does REU impose a norm of global uniformity on risk-sensitivity? Inappropriate because what it revises, [EUT], does the same thing, even more stringently (allowing only one trivial r). Still, once we attend to the requirement [i.e., the requirement that your risk-attitudes are fixed independently of stakes] it is natural to ask, why is it a norm of rational preference that an agent's risk-sensitivity is globally fixed in this way?' (Armendt 2014, p. 1123)

This is right: it is reasonable to ask why your low-stakes preferences settle your risk-attitudes when considering a coin toss on millions of dollars. (An aside: I do not think, as Armendt suggests, that it is inappropriate to ask why REU is stakes-insensitive on the ground that it revises EUT. WLU *also* revises EUT and yet is stakes-sensitive; stakes-insensitivity is an accidental feature of a constant w.)

Formally, we can characterise stakes-sensitivity as follows. Let l be a lottery with outcomes o_i and for $k > 0$, let kl be a lottery where each outcome in l replaced by ko_i, which has utility $k \cdot u(o_i)$. A theory is stakes-insensitive if it says for all acts l, m and positive k:

$$l \succcurlyeq m \text{ if and only if } kl \succcurlyeq km^{69}$$

[69] This is 'scale-invariance'. Another kind of stakes-insensitivity would be *shift-invariance*. Let $l + b$ be a lottery with each outcome o_i in l replaced by one of utility $u(o_i) + b$. A theory is shift-invariant if it satisfies for all b: $l \succcurlyeq m$ if

A stakes-insensitive theory says that your preferences are blind to the magnitude (in utilities) of the possible gains and losses you might face. Such a theory might pay attention to certain structural features of lotteries—the order of the outcomes and the ratios of utility-differences between them. But it does not pay attention to how bad the worst outcome is, nor the magnitude of the utility-differences between outcomes.

REU and EU are both stakes-insensitive.[70] WLU, however, is not. To illustrate, consider again $u(\$x) = \sqrt{x}$ and a scaled-down Allais case (each utility is scaled down by a factor of $1{,}000$ from the initial Allais case):

	Tickets 1-89	Ticket 90	Tickets 91-100
a^*	$\$1$	$\$1$	$\$1$
b^*	$\$1$	$\$0$	$\$5$

	Tickets 1-89	Ticket 90	Tickets 91-100
c^*	$\$0$	$\$1$	$\$1$
d^*	$\$0$	$\$0$	$\$5$

Our WLU-maximiser breaks with their high-stakes Allais preferences:

	$u(x)$	$w(x)$	$p_a(x)$	$p_b(x)$	$p_c(x)$	$p_d(x)$
$\$0$	0	1	0	$.01$	$.89$	$.9$
$\$1$	1	$.5$	1	$.89$	$.11$	0
$\$5$	2.236	$.400$	0	$.1$	0	$.1$

$$WLU(a^*) = 1$$

$$WLU(b^*) \approx 1.080$$

$$WLU(c^*) \approx 0.058$$

$$WLU(d^*) \approx 0.095$$

and only if $l + b \succcurlyeq m + b$. REU is shift-invariant while WLU is not, and most of what I say below applies to shift-invariance.

[70] This follows by noting that for all k, l, $EU(kl) = k \cdot EU(l)$ and $REU(kl) = k \cdot REU(l)$.

So $b^* \succ a^*$ and $d^* \succ c^*$, even though these are scaled-down versions of our original a, b, c, d.

WLU allows for stakes-sensitivity in a natural way. Because w takes outcomes as its domain, how you structure the potential realisation of your ends depends on what those ends are. We are risk-averse when buying parachutes not simply because death is worse than a broken leg, but because death is particularly *bad*. So, you might think that minimising the probability of death is important in a way that minimising the probability of the worst outcome is not when, say, buying vegetables.

Buchak might respond that stakes-sensitivity is unnatural because a comparison between l and m involves the same structural features as a comparison between kl and km (indeed, this point is made by Wilkinson Forthcoming, Section 5). But this raises the question of what structural features are and why we should care about them. If by 'structural features' you mean the ratio of utility-differences between outcomes, then it is true that l and m are structurally identical to kl and km. But it is unclear why we should care only about *these* structural features. For example, Hájek claims:

> '[T]is not contrary to reason to be more risk-averse when the stakes are high, but less risk-averse or even risk-neutral when the stakes are low. It's not even unreasonable, not normatively defective at all, I contend. I think it's no accident that Allais' example involves very high stakes. An example with the same sort of structure but tiny amounts of money would not be nearly as compelling.' (Hájek Forthcoming[a], Section 6)

Building on Hájek's point, I claim that it is not contrary to reason to consider properties that go beyond those fixed by scale-transformations. After all, the WLU theorist thinks that there is an important structural difference between l, m and kl, km: the average weight of each outcome is different after re-scaling utilities by k, and average weight is what determines how each outcome contributes to the instrumental utility of the lottery. Though the ratio of utility-differences remains fixed when multiplying utilities by k, the significance of, say, the worst outcome relative to the best might change. So, because you structure the potential realisation of your ends by relative weight, multiplying utilities by a constant can affect how you structure the potential realisation of your ends. Normatively significant features of a lottery do not, therefore, need not be stakes-insensitive.

Another reason to uphold stakes-insensitivity comes from Hájek (2014a pp. 9-10; Easwaran 2014a makes a similar point, as does Wilkinson Forthcoming). Since utility functions are unique only up to positive affine transformations, we should treat l and kl equivalently. After all, the fact that each outcome is assigned utility $u(o)$ and not $k \cdot u(o)$ is just down to our choice of representation. And choice of representation ought not change the ranking of lotteries.[71]

But WLU shows that this argument rests on an equivocation. WLU *is* unchanged on rescaling the utility function, since for acts $f = \alpha_1 o_1 + \cdots + \alpha_n o_n$ and $g = \beta_1 o_1 + \cdots + \beta_n o_n$, the following inequalities are equivalent for $k > 0$:

$$\sum_i \frac{w(o_i)}{\sum_j w(o_j)\alpha_j}\alpha_i \cdot u(o_i) \geq \sum_i \frac{w(o_i)}{\sum_j w(o_j)\beta_j}\beta_i \cdot u(o_i)$$

And

$$\sum_i \frac{w(o_i)}{\sum_j w(o_j)\alpha_j}\alpha_i \cdot (k \cdot u(o_i)) \geq \sum_i \frac{w(o_i)}{\sum_j w(o_j)\beta_j}\beta_i \cdot (k \cdot u(o_i))$$

That is, scaling the utility function by k does not change the WLU-ordering. But rescaling the utility function is *not* the same as changing the outcomes such that, for a given utility function, each is replaced by one of k-times its original utility. The former involves a change in representation (but not a change in outcomes), while the latter involves no change in representation (but a change in outcomes). And we have already seen that it is false that:

$$\sum_i \frac{w(o_i)}{\sum_j w(o_j)\alpha_j}\alpha_i u(o_i) \geq \sum_i \frac{w(o_i)}{\sum_j w(o_j)\alpha_j}\alpha_i u(o_i)$$

if and only if,

$$\sum_i \frac{w(ko_i)}{\sum_j w(ko_j)\alpha_j}\alpha_i u(ko_i) \geq \sum_i \frac{w(o_i)}{\sum_j w(o_j)\alpha_j}\alpha_i u(ko_i)$$

So, there is an important difference between scaling the *utility function* and scaling the *outcomes*. In the EU and REU frameworks we can conflate the two, but that is because *those* theories attend to a narrow kind of structural feature. WLU's richer structure gives us the language to carve between changes of stakes and changes of representation. Absent a good argument for REU or EU, we should not infer stakes-insensitivity from the re-scalability of the utility function.

[71] A clarification: Hájek (2014a) deals with the problem of how to assign instrumental utilities to infinite acts (those that assign non-zero probability to an infinite number of outcomes). As such, he might be best interpreted as provisionally assuming EUT as the correct rule for evaluating finite acts and then reasoning inductively—*since* our decision theory assumes stakes-insensitivity in the finite case, we should expect stakes-insensitivity to hold in the infinite case. This is not the argument in Wilkinson (Forthcoming), who infers directly from the re-scalability of the utility function to stakes-insensitivity for all lotteries.

1.7 Implications of Betweenness

To conclude, I want to sketch two properties of Betweenness-satisfying theories (properties that Betweenness-violating ones lack). Though some might take these as independent normative reasons to uphold Betweenness, it is not entirely clear to me that they have much appeal if we are antecedently suspicious of Betweenness. Nonetheless, they are both pre-theoretically compelling, so it is good to know that Betweenness vindicates some plausible features of our ordinary practical reasoning.

1.7.1 Information and Randomisation

Betweenness vindicates the following thought: we should never pay simply to make the world less predictable. Put slightly differently, losing information is never beneficial in and of itself.

Now, the Allais Preferences might require that you pay to avoid *free information*—simply recall in the Dynamic Allais case from the Introduction that you pay a small fee *not* to find out which of Tickets 1-89 or 90-100 were drawn. (If you find out which ticket was drawn, you end up making choices that result in a lottery that is, by your current lights, worse than one you can have outright.)

That is not a decisive reason to reject the Allais Preferences (cf. Buchak 2013, Section 6.5). The defender of risk-sensitivity will deny that there is a single privileged epistemic standpoint that controls your decisions. Employing a game-theoretic analogy, Machina (1989, footnote 26—he attributes the point to Karni) thinks of paying to avoid free information as a form of 'precommitment on the part of the "first-stage agents" against subsequent moves on the part of the "latter-stage agents". So, we do not turn down information because we *dislike* information. Rather, by turning down free information we give ourselves access to certain coordinating strategies between our time-slices, which enables us to choose what is, by our current lights, the best means to our ends.

But even if we pay to avoid information as a kind of future-self management, we might maintain (or hope it turns out to be the case) that information in and of itself is neither good nor bad. In particular, it is a pre-theoretically compelling idea that more (or less) information is only good (bad) insofar as it helps us achieve our goals.

Fortunately, Betweenness vindicates this intuition, or at least something very close to it. Betweenness-violators may pay a fee *simply* to randomise and end up with a less accurate map of the world. Consider:

> *Two Boxes*: In front of you is a box containing Ticket A and a box containing Ticket B. For a small fee, I can close the boxes and shuffle them such that you do not know which ticket you receive on selecting a box. What should you do?

It looks like the Betweenness-violator is paying to *lose* information here. There is some fact that they knew (where the tickets were), and they have paid to make it the case that they no longer know that fact. There are no future decisions to manage here, so we cannot think of this as a future-self management strategy. By paying to randomise, our Betweenness-violator does not give themselves access to different lottery tickets, coordinating strategies, or anything like that. It seems rather that they pay simply to get a less accurate map of the world.

Betweenness, however, never says that you should pay to randomise simply for the sake of it. This means that we vindicate the intuition that information itself is neither here nor there—we should never pay to randomise, nor should we pay to avoid randomising. So, while risk-sensitivity is closely tied up with aversion to certain kinds of information-aversion, Betweenness vindicates and explains the commonsense thought that you should never pay simply make the world more unpredictable.

1.7.2 Plans and Randomisation

A second important property of Betweenness is that it is equivalent to a natural constraint on plans. Gul and Lantto (1990, Theorem 2.ii) show that Betweenness (along with preferences being Complete, Continuous, Transitive, and respecting FOSD) is equivalent to:

> **Dynamic Program Solvability (DPS):** In a sequential decision, say that at some chance node nature might make one of two moves, E or $\neg E$. If you are permitted to choose Plan A and permitted to choose Plan B, then you are permitted to choose a plan that agrees with A in the event E and B in the event $\neg E$.

DPS says that rational plans can be mixed and matched. This can be thought of as a kind of weak counterfactual separability: when considering what to do in the event that there is a pandemic next year, you need not consider what rational course of action you would have chosen were there *no* pandemic (and vice versa). In other words, when making contingency plans you can ‘bracket off’ what you would have done in other scenarios, provided you trust that you would acted rationally in those scenarios.

Gul and Lantto (p. 173) call this a normative reason to accept Betweenness. It is, however, a little unclear what this amounts to. Consider the following simple case:

> *Dynamic Weather*: Tomorrow you can take Gamble A in the case of rain or shine. Similarly, tomorrow you can take Gamble B in the case of rain or shine. What about the plan ‘Gamble A if it rains, Gamble B if it shines’?

Since you do not know what the weather will be like tomorrow, ‘mixing plans’ in this way effectively counts as randomising—you give yourself some probability of getting Gamble A and some probability of getting Gamble B. Gul and Lantto think that violating DPS here is unmotivated, and this gives us reason to accept Betweenness. But in Dynamic Weather somebody like Buchak who rejects Betweenness might simply think, say, that the ‘mixed plan’ is impermissible because it has different global properties to the unmixed plans. So, I do not think that pointing to DPS by itself counts as an argument for Betweenness. Rather, I think we should say that Betweenness explains why DPS holds as a rational constraint on plans.

So again, Betweenness allows us to retain a key piece of ordinary practical reasoning—that we can make contingency plans separately, rather than having to simultaneously consider our behaviour in every possible contingency. Violations of DPS are indeed unnatural, and reasoning without DPS is demanding. So, Betweenness vindicates and important and natural way of thinking about plans.

1.8 <u>Conclusion</u>

I began with the idea that there is *something* plausible about Savage’s Businessman case—if I prefer one option to another in either of two cases, then I prefer it *simpliciter*. Decomposability, which is equivalent to Betweenness in the current setting, is a natural way of spelling that idea out.

I have argued that Betweenness-satisfying theories are preferable to Betweenness-violating ones, which includes REU, a popular and well worked-out alternative to orthodox EUT. Not only can we give WLU a plausible normative interpretation, it also vindicates a range of intuitive verdicts and features of our ordinary practical reasoning. So, while risk-sensitivity forces us to clarify the precise details of dominance, we can uphold reasonable risk-aversion *and* a plausible Savage-style commitment to dominance. Good things really do come to those who weight (in line with WLU).

1.9 Appendix: Proofs

Theorem 1: If u and $u \cdot w$ are increasing and w is decreasing, then WLU satisfies First-Order Stochastic Dominance.

Proof: Recall that we work with finite acts—those that assign non-zero probability to a finite number of outcomes.

For acts a, b, if a First-Order Stochastically dominates b, we can obtain b from a with a sequence of probability-shifts from each outcome to the next worst. For example, let $o_1 > o_2 \ldots > o_n$ and take $a = p_a(o_1) \cdot o_1 + \cdots + p_a(o_n) \cdot o_n$ (some probabilities might be 0). Without loss of generality consider a δ-probability shift ($\delta > 0$) from the best to the next-worst outcome, that is $a_\delta = (p_a(o_1) - \delta) \cdot o_1 \; + (p_a(o_2) + \delta) \cdot o_2 + \cdots + p_a(o_n) \cdot o_n$. It suffices to show $WLU(a) \succ WLU(a_\delta)$. Now:

$$WLU(a) = \frac{p_a(o_1)w(o_1)u(o_1) + p_a(o_2)w(o_2)u(o_2) + \cdots + p_a(o_n)w(o_n)u(o_n)}{p_a(o_1)w(o_1) + p_a(o_2)w(o_2) + \cdots + p_a(o_n)w(o_n)}$$

$$WLU(a_\delta) = \frac{(p_a(o_1) - \delta)w(o_1)u(o_1) + (p_a(o_2) + \delta)w(o_2)u(o_2) + \cdots + p_a(o_n)w(o_n)u(o_n)}{(p_a(o_1) - \delta)w(o_1) + (p_a(o_2) + \delta)w(o_2) + \cdots + p_a(o_n)w(o_n)}$$

Assuming u is increasing and w is decreasing, call the denominators D_a and D_{a_δ} respectively. Then $D_a - D_{a_\delta} = -\delta\big(w(o_2) - w(o_1)\big) = -\delta K$, where $K = w(o_2) - w(o_1) > 0$ since w is decreasing. Then:

$$WLU(a) - WLU(a_\delta) = \frac{(D_a + \delta K)[\sum_i p_a(o_i)w(o_i)u(o_i)] - D_a(\sum_i p_{a_\delta}(o_i)w(o_i)u(o_i))}{D_a(D_a + \delta K)}$$

Since $\delta K, D_a > 0$, it suffices to show the numerator is positive. On cancelling terms, this amounts to showing $u(o_1)w(o_1) - u(o_2)w(o_2) > 0$, which since $o_2 < o_1$ holds if $u \cdot w$ is increasing. □

Theorem 2: If u is increasing, w is decreasing and $k(x,y) = \frac{u(x)-u(y)}{w(x)-w(y)}$ is decreasing in both x and y, then WLU satisfies Second-Order Stochastic Dominance.

Proof: Recall that we work with finite acts. For acts a and b, if a Second-Order Stochastically Dominates b then b results from a sequence of transformations which shift pairs of probabilities either side of the mean farther away, leaving the mean unchanged. That is, beginning with a and letting $u(o_i) < u(o_{i+1}) < EU(a)$ and $u\left(o_{j+1}\right) > u\left(o_j\right) > EU(a)$, we make transformations of the form:

$$a = p_a(o_1) \cdot o_1 + \cdots p_a(o_n) \cdot o_n$$

To:

$$a_{\delta,\epsilon} = a + (\delta \cdot o_i - \delta \cdot o_{i+1} - \epsilon \cdot o_j + \epsilon \cdot o_{j+1})$$

Assume that the utility-scale is chosen such that each $u(o_i) > 0$.

Crucially, since $a_{\delta,\epsilon}$ is mean-preserving, $EU(a) = EU(a_{\delta,\epsilon})$, meaning:

$$\delta \cdot u(o_i) - \delta \cdot u(o_{i+1}) - \epsilon \cdot u(o_j) + \epsilon \cdot u(o_{j+1}) = 0$$

I refer to this as the *mean-preserving equality*. It suffices to show that $WLU(a) - WLU\left(a_{\delta,\epsilon}\right) > 0$, that is:

$$\sum_l \frac{w(o_l)}{\sum_m p_a(o_m)w(o_m)} p_a(o_l)u(o_l) - \sum_l \frac{w(o_l)}{\sum_m p_{a_{\delta,\epsilon}}(o_m)w(o_m)} p_{a_{\delta,\epsilon}}(o_l)u(o_l) > 0$$

Which is:

$$\sum_l \frac{w(o_l)(((\sum_m p_{a_{\delta,\epsilon}}(o_m)w(o_m))u(o_l)p_a(o_l)) - (\sum_m p_a(o_m)w(o_m))u(o_l)p_{a_{\delta,\epsilon}}(o_l)u(o_l))}{R}$$

$$> 0$$

Where R is the denominator obtained by cross-multiplying. Since R and $w(\cdot)$ are positive, it suffices that:

$$\sum_l (\sum_m p_{a_{\delta,\epsilon}}(o_m)w(o_m))u(o_l)p_a(o_l) - (\sum_m p_a(o_m)w(o_m))u(o_l)p_{a_{\delta,\epsilon}}(o_l)u(o_l)) > 0$$

Expanding this gives:

$$\sum_{l}[(\sum_{m} p_a(o_m)w(o_m))p_a(o_l)u(o_l)$$

$$+\left(\delta w(o_i) - \delta w(o_{i+1}) - \epsilon w\big(o_j\big) + \epsilon w\big(o_{j+1}\big)\right) p_a(o_l)u(o_l)]$$

$$-\sum_{l}[(\sum_{m} p_a(o_m)w(o_m))p_a(o_l)u(o_l) +$$

$$(\sum_{m} p_a(o_m)w(o_m))(\delta u(o_i) - \delta u(o_{i+1}) - \epsilon u\big(o_j\big) + \epsilon u(o_{j+1}))]$$

This simplifies quickly. The last line is 0 by the mean-preserving equality, and the first and third lines cancel, leaving us to show of the second line:

$$\sum_{l}\left(\delta w(o_i) - \delta w(o_{i+1}) - \epsilon w\big(o_j\big) + \epsilon w\big(o_{j+1}\big)\right) p_a(o_l)u(o_l) > 0$$

Note also that each $p_a(o_l)u(o_l)$ is positive, so it suffices to show:

$$\delta w(o_i) - \delta w(o_{i+1}) - \epsilon w\big(o_j\big) + \epsilon w\big(o_{j+1}\big) > 0$$

Again, by the mean-preserving equality we know:

$$\delta = -\epsilon \frac{u\big(o_j\big) - u\big(o_{j+1}\big)}{u(o_i) - u(o_{i+1})}$$

Substituting this in and cancelling gives:

$$\frac{u(o_{i+1}) - u(o_i)}{w(o_{i+1}) - w(o_i)} > \frac{u\big(o_{j+1}\big) - u\big(o_j\big)}{w\big(o_{j+1}\big) - w\big(o_j\big)}$$

And since $o_i < o_j$, this holds if $k(x, y) = \frac{u(x)-u(y)}{w(x)-w(y)}$ is decreasing in both arguments. □

Chapter 2

Stalemate: Dominance, Information, and Instability

2.0 States, Causation, and Evidence

I now turn to the first dominance principle discussed in the Introduction: *State-wise Dominance*. Surely if one act does better however the world turns out to be, then it does better *simpliciter*—you simply have no reason to perform an act that will turn out worse. Though this idea is intuitively plausible, it is no easy matter to say precisely what State-wise Dominance amounts to. In this and the next two chapters, I investigate State-wise Dominance through the lens of *Causal Decision Theory* and the (in)famous Newcomb Paradox. My conclusion will be that a widely accepted version of State-wise Dominance, Causal State-wise Dominance, is in trouble. The goal of this chapter is to demonstrate that while Causal State-wise Dominance represents one way of spelling out the intuition behind State-wise Dominance, a rival evidential formulation does just as well. This lays the foundation for arguing (after a detour through the subtleties that arise for the causal view in deterministic cases in Chapter 3) that the evidential standard has strong pragmatic advantages in Chapter 4. Though I will not decisively reject Causal State-wise Dominance, these chapters chronicle my journey from being a contented causalist to a conflicted evidentialist.

To begin, consider the following scenario:

> *Job Market*: You are considering whether or not to prepare for your job interview tomorrow. You reason that if you get the job, then preparation is a waste of time (you could go to the pub instead), and similarly that if you do not get the job, then preparation is also a waste of time (you could go to the pub instead). We might represent this:

	Get Job	Don't Get Job
Prepare	10	0
Don't Prepare	15	5

It seems that not preparing is dominated by preparing: however the world turns out to be (i.e., whether you get the job or not), you prefer not to prepare. But surely you make a mistake by not preparing! Your reasoning goes wrong in Job Market because what you choose affects whether you get the job or not. So, you ought not reason about what you prefer *holding fixed* whether you get the job, since whether you get the job is *not fixed* independently of your (lack of) preparation.

Stepping back into the formal language of decision theory, recall that I defined acts as mappings from states to acts. While it is up to you which act you choose, states represent the parts of the world that are settled independently of what you do. Job Market shows just why this matters: if the states are not settled independently of what you do, you will end up in the pub far more often than you ought. And this is where we meet our first challenge to State-wise Dominance: what precisely does it mean for a state to hold *independently* of what you do?

A widely shared thought is that states must be *causally independent* of your decision. It would be futile to deliberate with the goal of changing things outside your causal influence—say the past or near-future events on faraway planets. So, causalists think that states describe what it outside of your control, which they analyse as 'what is causally independent of your decision'. They therefore reject your dominance reasoning in Job Market because the states—'Get Job' and 'Don't Get Job'—are not causally independent of your decision. We cannot infer that one act is preferable because it is preferable in either of those cases.

Lewis (1981a, pp. 11-14) gives an analysis of states that clarifies the kinds of thing can count as states. Lewis calls a *dependency hypothesis* a proposition that (i) is causally independent of your decision, and (ii) describes the full pattern of causal dependencies between acts and outcomes. I will follow Lewis in using k to refer to a dependency hypothesis. In order to specify the contents of each dependency hypothesis, we have to say what causal independence is. If, say, we adopt a counterfactual analysis of causal dependence, then 'the dependency hypotheses being causally independent of your decision' amounts to the requirement that for each k and act a_i:[72]

$$k \text{ if and only if } a_i \rightarrow k$$

Here, the counterfactual connective → is given a non-backtracking interpretation (or whichever species of counterfactual appears in our best analysis of causal dependence). Alternatively, you might analyse each dependency hypotheses in terms of chance functions (cf. Skyrms 1982, 1984), causal graphs (cf. Stern 2017), factors causally upstream of your acts (cf. Gallow 2020), or whatever notion you think appropriate. For the purposes of this chapter, I assume that counterfactuals have something to do with causation and decision; in particular, I will assume that if some dependency hypothesis holds, it would hold if you decided to do otherwise. A natural proposal then is that each k is a conjunction of act-outcome (non-backtracking)

[72] See Gibbard (1986) for this characterisation of states.

counterfactuals: each k is a hypothesis about which outcome would result from each of your acts.[73]

Recall that in the Savage-framework each act maps each state to a unique outcome. If we let each k be a proposition of the form $(a_1 \rightarrow o_1)$ & ... & $(a_n \rightarrow o_n)$, we can now describe each act by stating, for each dependency hypothesis, which outcome would occur by the lights of that dependency hypothesis. This provides the requisite Savage-mapping from states to outcomes. For ease of notation, let $o_{a,k}$ be the outcome such that doing a in state k would result in that outcome (we can think of $o_{a,k}$ as $a(k)$ in Savage's terminology; note that for distinct k and k', $o_{a,k}$ and $o_{a,k'}$ need not be distinct). Letting K be a full set of dependency hypotheses:

> **Causal State-wise Dominance:** If for each $k \in K$, $o_{f,k} \geq o_{g,k}$, then $f \succcurlyeq g$. Moreover, if for some $k \in K$, $o_{f,k} > o_{g,k}$, then $f \succ g$.

Since the states in Job Market are not causally independent of what you do, they are not dependency hypotheses, and Causal State-wise Dominance does not recommend that you go to the pub.

A paradigmatic theory that respects Causal State-wise Dominance is *Causal Expected Utility Theory*. If your preferences over outcomes can be represented with a utility function unique up to positive affine transformation, then the *causal expected utility* of act f is:

$$U_C(f) = \sum_k C(k) \cdot u(o_{f,k})$$

And Causal Expected Utility Theory says that you may choose a from option set A just in case a maximises U_C in A.[74] (A terminological note: some refer to Causal Expected Utility Theory simply as Causal Decision Theory. This conflates the question of whether to define states causally with the question of whether rationality requires EU-maximisation. I will use Causal Decision Theory (CDT) to refer to any decision theory that endorses Causal State-wise Dominance. Most of the time, the distinction between EU and non-EU versions of CDT will not matter, so I only refer to non-EU versions of CDT when the distinction matters.)

[73] I am assuming a fact of the matter (a so-called counterfact) about which outcome each act would bring about in each dependency hypothesis—no chance distributions over outcomes. This is a modelling assumption that might well be challenged (see Hájek Forthcoming[b] for survey and criticism of counterfacts).

[74] Note that I am assuming the existence of a utility function—a measure of desirability for each outcome. If we interpret dependency hypotheses as Savage-states, we can utilise Savage's representation theorem. Or, we might work with a primitive utility function and set aside questions of representation here.

It is common and often convenient to state U_C in slightly different terms. If each k is a list of counterfactuals, then we know that k holds if and only if $a \rightarrow o_{a,k}$ holds for each act a, so we can re-write:

$$U_C(a) = \sum_k C(a \rightarrow o_{a,k}) \cdot u(o_{a,k})$$

Causal State-wise Dominance leads to a contested verdict in the following case (from Nozick 1969):

> *Newcomb's Paradox*: In front of you are two transparent boxes. One box is transparent, and you can see that it contains $\$1{,}000$. The other box is opaque, and it contains either $\$1{,}000{,}000$ or $\$0$. The contents of the opaque box were determined yesterday by an uncannily accurate predictor (they get things right 90% of the time). If they predicted that you take just the opaque box ('one-boxing'), they placed $\$1{,}000{,}000$ in that box. If they predicted that you take both boxes ('two-boxing'), they placed $\$0$ in that box:

	Million	No Million
One-box	$\$1{,}000{,}000$	$\$0$
Two-box	$\$1{,}001{,}000$	$\$1{,}000$

Causal State-wise Dominance requires that you two-box. Since the predictor made their decision yesterday, the contents of the boxes are not something you can causally influence. Therefore, 'Million' and 'No million' are causally independent of your choice and fully describe act-outcome dependencies.[75] They are therefore dependency hypotheses, and Causal State-wise Dominance says that you should two-box.

Irrational, many say![76] While two-boxing does better holding fixed facts outside your causal influence, it is not obvious that two-boxing is a better means to your ends than one-boxing. After all, one-boxing increases the probability of becoming a millionaire, and who doesn't want to be a millionaire?

Horgan (1981) notes that there are non-causal senses of dependence that might be important in decision situations like Newcomb's Paradox. There is a reasonable sense in which the following claim is likely true: 'if I were to one-box, the predictor would have guessed it and there would be

[75] Note that this case is highly simplified: it assumes that there are only two ways that uncertainty could be resolved, and that money is all you care about.

[76] Prominent defences of one-boxing include Horgan (1981), Price (1986), and Ahmed (2014b).

a million in the opaque box; if I were to two-box, the predictor would have guessed it and there would be nothing in the opaque box'. These are a kind of *backtracking* counterfactual, specifying how the world would have been for you to perform each act. Perhaps it is these judgements of dependence that matter. Or we might reject the relevance of counterfactuals altogether. For example, Bradley (2017) thinks that what matters is the probability of *indicative* conditionals like 'if I one-box, I am a millionaire' and 'if I two-box, I am poor'. Or perhaps what matters is just what is most likely to bring about the good regardless of which conditionals hold, and one-boxing is clearly more likely to bring about the good than two-boxing.

I will assume that conditionals matter for decision theory. I have no decisive argument for this assumption, though it seems intuitive that deliberation, difference-making, and conditionals have *something* to do with each other. Moreover, in order to make any headway in Newcomb's Paradox, I think we had best keep the causalist and the evidentialist on the same page for as long as possible—framing the debate as one about which conditionals matter is one way of doing so. Following Horgan then, we can note that counterfactuals can be precisified in many ways. Our ordinary use of counterfactuals seems to be non-backtracking by default (cf. Gibbard and Harper 1978, Lewis 1979, Hedden Manuscript), and this might lend *prima facie* support to causalism. But in cases where your acts provide evidence for the states, Horgan (1981, p. 246) claims that the *pragmatically appropriate* reading of counterfactuals is always such that your credences satisfy the constraint that for all acts a and states k:[77]

$$C(a \rightarrow k) = C(k|a)$$

Since I have been working with act-outcome rather than act-state dependencies, we might tweak Horgan's view and say that the pragmatically appropriate reading of counterfactuals satisfies, for each act a and outcome o:

$$C\left(a \rightarrow o_{a,k}\right) = C(o_{a,k}|a)$$

Note that Bradley (2017, Chapters 6 and 7) constructs an indicative conditional that meets this constraint more generally: for any propositions x and y, Bradley's conditional ⇒ satisfies $C(x \Rightarrow y) = C(y|x)$.[78] Now, Horgan's backtracking counterfactual is not a Bradley-conditional (Horgan's equality holds only for act-state or act-outcome conditionals and does not satisfy the equality $C(x \rightarrow y) = C(y|x)$ for all propositions x and y; see Horgan 1981, p. 346).

[77] Note that Horgan (1981, p. 346) does not endorse the following equality for all *propositions*—that the probability of the conditional is a conditional probability holds only for act-state conditionals.

[78] See Bradley (2017) Chapter 8 for a discussion of how this conditional evades various triviality results (though see Hájek Forthcoming[b] for criticism).

Nonetheless, for conditionals that matter for decision theory—act-outcome conditionals—the two interact with probability in the same way. So, in the spirit of Horgan, I suggest that we let each dependency hypothesis be a conjunction of Bradley conditionals such that each conditional specifies which outcome is brought about by each act (and we can remain neutral on whether Bradley-conditionals correspond to ordinary English indicative conditionals, backtracking subjunctives, or some other flavour of conditional).[79] This lets us define a new quantity:

$$U_E(a) = \sum_k C(k) \cdot u(o_{a,k})$$

Each k is a conjunction of Bradley-conditionals that specify which outcome occurs if you perform each act—each k is of the form $a_1 \Rightarrow o_1 \wedge \ldots \wedge a_n \Rightarrow o_n$. And since this conditional tracks statistical connections between acts and outcomes, $C(a \Rightarrow o) = C(o|a)$, and noting that k holds if and only if $a \Rightarrow o_{a,k}$, we can re-write:

$$U_E(a) = \sum_k C(o_{a,k}|a) \cdot u\left(o_{a,k}\right)$$

$U_E(a)$ is the *evidential expected utility* of a—relevant act-outcome dependences always tracks track act-outcome statistical dependencies. And *evidential decision theory* (EDT) says that you should maximise U_E.

EDT recommends one-boxing—$U_E(\text{One} - \text{box}) = \$900{,}000$ while $U_E(\text{Two} - \text{box}) = \$101{,}000$. On EDT, the rational thing to do is increase the probability of the good, regardless of whether you causally promote it.

Irrational, many say! On the standard reading of counterfactuals, what is in the opaque box is independent of what you do now—its contents were settled yesterday. And yet EDT says that you should evaluate two-boxing assuming the predictor placed nothing in the opaque box, while simultaneously evaluating one-boxing assuming the predictor placed a million in the opaque box. This is to assess different acts relative to different pasts, which makes it sound a lot like you can affect the past (and you cannot).

[79] All that really matters here is that there is a kind of conditional that tracks statistical dependencies between acts and outcomes. Reasoning on the basis of such conditionals I will variously refer to as 'backwards looking', 'backtracking' and 'evidential' reasoning. If you object strongly to my usage of any of those terms, feel free to substitute another. For example, you might take issue with my use of the term 'backtracking', since outcomes may occur *after* acts (and perhaps you want to reserve 'backtracking counterfactuals' for those whose consequent occurs *before* its antecedent). I use the term backtracking to highlight the salient feature of these conditionals (the feature that the causalist rejects): that act-outcome dependencies reflect how the past would likely have been for you to perform each act.

In response, the evidentialist can accept that the standard reading of counterfactuals is non-backtracking but point out that Newcomb's Paradox is not a standard case. Who is to say that the standard reading of counterfactuals dictates what is pragmatically appropriate in bizarre circumstances?[80] Furthermore, they can accept that we cannot change the past—there are facts outside your causal influence and that what you do makes no difference to those facts.

The evidentialist simply denies that what matters for the purpose of practical deliberation is what holds *causally* independently of your choice. It may be pragmatically appropriate to reason about each act relative to the past that makes the act likely, but this does not presuppose a belief that you can change the past. We must distinguish between *assessing* each act relative to different past conditions and acting *in order* to change the past.[81] Moreover, Ahmed (2014b, p. 143) and Horgan (2017, pp. 40-44) have articulated a good reason to care about backtracking conditionals (or indicative conditionals, or Bayesian conditional probabilities). By paying attention to statistical correlations, you treat yourself 'as a part of nature'. By thinking in non-backtracking terms, as the causalist does, you hold fixed the predictor's decision and so take worlds in which you beat the predictor just as seriously as worlds in which you do not. But you do not ascribe to yourself superpowers or *sui generis* abilities that let you circumvent the predictor's accuracy! So, you should take seriously facts about what the world would have to have been like in order for you to behave in various ways. Rather than intervening *ex nihilo* and ignoring statistical correlations, the evidentialist thinks about themselves as embedded in the world and so pays attention to what the past would have to have been like for them to act differently. If I two-box, the predictor likely guessed it—this is the conditional that matters for the purposes of evaluating acts. If the predictor guessed that I two-box, I would do better by one-boxing—this makes it sound like I can intervene to beat the predictor, so it is not the conditional that matters for the purposes of evaluating acts.

The debate over CDT and EDT is sharp and has gone through many phases.[82] Personally, I have gone through phases of thinking that two-boxing is *obviously* the correct thing to do, and I have

[80] Moreover, sophisticated defences of EDT (see Eells 1982 and Price 1986) show that EDT coincides with CDT in cases where our intuitions are most strongly favour the causalist.

[81] Cf. Lewis' (1981a, p. 9) comment that the evidentialist is guilty of 'managing the news'. 'Managing the news' makes it sound like the EDT'er acts *in order* to keep themselves in the dark. But the U_E-maximiser need not place any intrinsic value on avoiding bad news, and U_E-maximisation does not say that your reasons for performing an act come your epistemic state on performing that act. The U_E-maximiser acts *in order* to bring about preferable outcomes—this is importantly different to acting *in order* to bring about a high credence in the good. Thanks to Toby Solomon for discussion here.

[82] The debate begins with Nozick (1969). Jeffrey (1983) systematically developed a view that became known as Evidential Decision Theory (Jeffrey at one stage rejected EDT; he ultimately concluded, 2004 p. 13, that Newcomb's Paradox is not a genuine decision). Early causalist views include Gibbard and Harper (1978), Skyrms (1982), Sobel (1988), and Lewis (1981a).

gone through phases of thinking that one-boxing is *obviously* the correct thing to do. Both views are underwritten by some intuitive account of conditionals in decision-making, and both accounts can be motivated by a coherent story. What we need are arguments for one view over the other, which I now examine.[83]

2.1 Against Evidentialism I: Full Information

Imagine a well-meaning bystander who has looked into the opaque box. While you do not know what the bystander sees, you do know that they will advise two-boxing *whatever* they see (Nozick 1969 originally discussed the case of a well-wishing friend). This is the argument from *Full Information*: someone who is identical to you, except that they know how your subjective uncertainty will be resolved, tells you to two-box. EDT conflicts with the plausible principle that you should always defer to someone who knows more than you (indeed, that you should always defer to your fully informed self).[84]

There is, however, a straightforward response to this argument on behalf of the evidentialist. When you act on the bystander's advice, you hold fixed what they tell you. For example, you might reason to yourself, 'If they saw nothing in the box, then I would be better off two-boxing than one-boxing'—this is to compare two-boxing to one-boxing *holding fixed* the contents of the box. Or you might reason to yourself, 'If they saw a million in the box, then I would be better off two-boxing than one-boxing'—again this is to compare two-boxing to one-boxing *holding fixed* the contents of the box.

But the Job Market case shows how careful we must be in what we hold fixed. You are only entitled to hold fixed what is independent of your choice, and the evidentialist already thinks that the bystander's advice is not independent of your choice. That is not to say that you can causally influence what the bystander sees. Rather, it is to emphasise the pragmatic importance of conditionals of the form, 'If I two-box, the predictor likely guessed it and the bystander will be seeing a million; if I one-box, the predictor likely guessed it and the bystander will be seeing

[83] A note on scope: the debate over CDT and EDT is too big. This chapter is not an attempt to systematically review every argument for both positions. Rather, I work through some standard arguments that have received recent attention and so are worth commenting on especially. Moreover, each argument I discuss illustrates a common theme: the initial disagreement between CDT'er and EDT'er makes it hard to give an argument for one view that does not talk past the other.

[84] I have already rejected that you must defer to someone who knows more than you (but does not have access to all the relevant facts) in the Allais case. Crucially, in Newcomb's Paradox the bystander knows which *state* holds, thus resolving all uncertainty.

nothing'. And there is nothing incoherent in acting on the basis of these conditionals while maintaining that you cannot causally affect what the bystander sees. Just as the contents of the box is not independent of what you do, the bystander's advice is not independent of what you do—thinking about a bystander merely pumps the intuition in favour of one standard of independence.

Maybe you find this perplexing. Skyrms (1984, p. 67) argues that the evidentialist is committed to turning down free information—you would pay not to know what the bystander sees. And you might think that free information can never be a bad thing.

But finding out what your friend sees is not free by the evidentialist's lights. After all, you know that you will two-box on seeing what the bystander sees (since on learning what is in the box, you will conditionalise on the contents of the box, and $U_E(\mathrm{Two-box}) > U_E(\mathrm{One-box})$ after conditionalising), so learning what the bystander sees is correlated with the predictor having placed nothing in the opaque box. The evidential expected utility of learning is therefore lower than that of not learning—we are in the unusual situation of a learning experience being correlated with some inauspicious state holding. Learning the contents of the opaque box is only 'free' in the informal sense of leaving your wallet no lighter. But on the pragmatically appropriate analysis of 'dependence', the contents of the box are dependent on your decision to learn about what is in the box. (Analogy: in Job Market, you know that if you conditionalise on either 'Get Job' or 'Not Get Job', you will prefer to go to the pub. And you will indeed conditionalise on one of these facts. But you currently think it a mistake to go to the pub because whether you do so or not affects what you will conditionalise on.)[85]

You might think that there is something like an epistemic duty to gather more information. If so, then Newcomb's Paradox may simply be a case in which pragmatic and epistemic norms conflict. But that happens all the time—consider research with safety risks—and the evidentialist will not distinguish between 'turning down free information because it is indicative of something bad' and 'turning down free information because it causes something bad'. On the other hand, if you follow Good (1967) in reducing the duty to gather information to an epistemic norm, then the evidentialist can simply maintain that Good-style arguments fail in Newcomb's Problem.

[85] Adams and Rosenkrantz (1980) discuss the value of information in Jeffrey's decision theory—they note that the expected value of information may be negative when states are not act-independent. They do not discuss Newcomb's Paradox directly, so I am unsure whether they intend their point to apply when there are non-causal act-state dependencies. Skyrms (p. 67) criticises the evidentialist by saying 'peeking [i.e., getting more information] won't change what's under the boxes'. But the evidentialist might accept that they cannot *change* what's under the boxes (in the sense of causally influence the contents of the box)—nonetheless, the contents of the boxes are *dependent* on your decision to peek or not.

Finding out the contents of the opaque box before making your decision decreases the probability of the good.

But still, you might imagine your friend's incredulous stare on hearing that the contents of the box in front of them, which has been sitting there since yesterday, depends on what you do now! Maybe you have this incredulous stare yourself. But we must be careful—when we think about dependence in the abstract, we likely default to the standard non-backtracking reading of dependence, rather than keeping the predictor's efficacy in the front of our minds. And the evidentialist will be quick to remind their friend that they are not positing miraculous powers that let them causally influence the past—they are merely denying that 'decision-making in Newcomb's [Paradox] should be based on causal efficacy' (Horgan 1981, p. 342).

2.2 Against Evidentialism II: Actual Value

But surely, responds the causalist, what matters is what is *actually* in the box. And on any reasonable sense of 'actual', the contents of the box are actually fixed—the bystander can see what is actual! This line of thought moves us from an argument from Full Information to one from *Actual Value*.

Plenty of philosophers think that actual value matters. Schoenfield (2014), as discussed in the Introduction, claims that anyone who cares about actual value would be engaged in 'expected [or instrumental] value fetishism' were they to act in a way that is certainly less valuable than an available alternative. Doody (2016) gets at a similar idea when he defines the actual value of an act as its utility in conjunction with the correct (causal) dependency hypothesis holding; he then argues that EDT is incompatible with an actual value conception of rationality (one on which the goal of rationality is to promote actual value).[86] Since the contents of the box is already fixed, two-boxing is actually more valuable than one-boxing, so the argument goes.

But again, the evidentialist has a ready-made reply.[87] Firstly, they can note that there are four possible scenarios, which we can think of as four possible worlds, in Newcomb's Paradox:

[86] Doody (2016, p. 12) simply defines the actual value of a to be the value of a conjoined with the actual causal dependency hypothesis holding. (In a similar spirit, Spencer and Wells 2019 define the actual value of a to be the value of the outcome o such that $a \rightarrow o$ holds in the actual world, where this counterfactual is a non-backtracking one.)

[87] After writing a precursor to this chapter in 2018, I learned that Ahmed & Spencer (2020) make a similar point in their 'Objective Value is Always Newcombizable', Section 4. I deal with their improved version of the Actual Value argument shortly.

w_1: One-box and a million
w_2: One-box and nothing
w_3: Two-box and a million-plus-thousand
w_4: Two-box and a thousand

When we say that two-boxing is *actually* more valuable than one-boxing, what are we saying? Not that each two-box world is more valuable than each one-box world—the world in which you one-box and become a millionaire is more valuable than the world in which you two-box and get a thousand. And clearly we are not making a claim about which *world* is actual, since we are comparing two different acts and no single world is both a one- and two-box world. Indeed, for it to make sense to deliberate about what to do, you cannot have settled which of $w_1 - w_4$ is actual. Rather, the claim seems to be that you know that the actual world must be one of two ways, either it is w_1-or-w_3 *or* it is w_2-or-w_4. That is, regardless of what you do, we hold the contents of the box fixed when talking about what is actual.

But here is where the evidentialist's ready-made reply becomes apparent—we are not entitled to hold fixed the contents of the opaque box when reasoning about what is actual! Say that the actual world is w_1. What then is the 'actual value' of two-boxing given that you one-box in the actual world? (Note the strangeness of this question might make us question the relevance of 'actual value' talk when applied to options in the first place.) If we hold fixed facts causally independent of your choice, then were you to do otherwise the contents of the opaque box would be the same and you would end up in w_3. But as before, the evidentialist thinks that causally independent facts are the wrong facts to hold fixed from a pragmatic perspective—if you were to do otherwise, the predictor would likely have guessed it and done otherwise. Actually, you cannot beat the predictor. So, the evidentialist will reject Doody's characterisation of an act's 'actual value' as 'value given (or in conjunction with) the correct causal dependency hypothesis'. They will similarly deny Spencer and Wells' (2019) characterisation of a's actual value as the value of o such that $a \rightarrow o$ holds (where $\rightarrow$ is non-backtracking). Defining actual value in that way presupposes that for pragmatic purposes what 'actually holds were I to do otherwise' is what holds in virtue of facts outside my causal influence—an unattractive view for anyone who thinks that deliberation should be based on evidential, not causal, reasoning. (I personally feel torn when talking about what 'actually' holds in Newcomb's Paradox. When I keep the predictor's efficacy in the front of my mind, I have no strong intuition that the actual contents of the box are fixed.)

Perhaps 'what actually holds' just *is* a causal concept—'*x* actually holds if *x* holds in virtue of facts outside my causal influence' might be something like an analytic truth. For example, Spencer and Wells (2019, p. 44) claim that it is uncontroversial that 'taking both boxes … uniquely maximises actual value'. I do not think so. It might be analytically true that what is actual is what holds in the actual world, rigidly designated. But again, the actual world (rigidly designated) is just a world in which you one-box or a world in which you two-box—to talk about what actually holds were you to do otherwise strikes me as up for grabs. And again, when I focus on the predictor's efficacy, I am at a loss to say what 'actually' holds in Newcomb's Paradox.

Ahmed and Spencer (2020) have recently developed a more sophisticated version of the argument from actual value. They claim that EDT is incompatible with acts having objective values (I take their use of 'objective value' to correspond to what I have been calling 'actual value'). Ahmed and Spencer really present two arguments for EDT's incompatibility with options having objective value—one from 'Expectationism' (2020, Sections 6-7) and one from 'Newcombizability' (2020, Section 8). They ultimately conclude that only the Newcombizablity argument succeeds. Nonetheless, I will focus on the argument from Expectationism since it relies on simpler premises, and the key premise that I reject in the argument from Expectationism is presupposed in their argument from Newcombizability (I highlight the relevant premise below). So, while I focus on the simpler argument for ease of exposition, the basic point applies to their more involved argument from Newcombizability.

The key premise in Ahmed and Spencer's argument is that actual value should be the kind of thing you can *learn* about by performing various acts. Formally, let $O(a) = v_i$ be the proposition that the objective value of a is v_i. Ahmed and Spencer (p. 1173) argue that a rational agent's credences may satisfy *Relevance*, which is made up of three sub-constraints:

i) $C\big((O(a) = v_1 \vee O(a) = v_2)\big) = 1$

ii) $C((O(a) = v_1)) = x$, for some $x < 1$,

iii) $C\big((O(a) = v_1)\big|a\big) = y$, for some $y \neq x$

Essentially, i) says that it is permissible to be certain that an option has one of two objective values, ii) says that you need not be certain which of those values it is, and iii) says that the option in question can provide evidence for its own objective value. iii) is the key premise, and I will refer to it as the *learnability premise*: whatever objective value is, it is the kind of thing you can learn about through your own acts. Ahmed and Spencer (p. 1172) take the learnability premise as a necessary pre-condition for value being *objective*, '[e]very remotely plausible conception of objective value must allow an agent to regard an option … as evidence about what the objective value [of that option] is'. (More on this shortly.)[88]

They then show that if an act's instrumental utility is the expectation of its objective value (i.e., if $U(a) = \sum_v C(O(a) = v) \cdot v)$, then i)-iii) are inconsistent with EDT. So, if i)-iii) really do characterise 'every remotely plausible conception of objective value', then EDT is incompatible with act's having objective values.[89]

Here is their argument as it applies in a concrete case (see Ahmed and Spencer 2020, Appendix A for a general version of their argument): say that $C(O(a_1) = 1) = C(O(a_1) = 0) = .5$ and that $C(O(a_2) = 1) = C(O(a_2) = 0) = .5$. That is, you are assign 50% probability to each of two options having objective value 1 and 0, respectively. But say that doing a_1 provides evidence for its own objective value, $C(O(a_1) = 1|a_1) = 1$, and that doing a_2 provides no evidence for any proposition about objective value. Now, the expected objective value of each act is:

$$EV\big(O(a_1)\big) = .5 \cdot 1 + .5 \cdot 0 = .5$$

$$EV\big(O(a_2)\big) = .5 \cdot 1 + .5 \cdot 0 = .5$$

And yet $U_E(a_1) = 1 > U_E(a_2) = .5$. So, EDT says you may not take a_2, though a_2 maximises expected objective value. More generally, when expected objective value diverges from U_E, EDT tells you not to maximise expected objective value but U_E instead. Objective value therefore does not play the right kind of action-guiding role.

[88] Their full argument from Newcombizability relies on what they call *Baseline Relevance* (p. 1178), which is that a rational agent's credences may satisfy for some $z > 0$:

i) $C(O(a_1) = v_1 \vee O(a_1) = v_3) = 1$
ii) $C(O(a_1) = v_1|a_1) = x < 1$
iii) $C(O(a_1) = O(a_2) + z) = 1$
iv) $C(O(a_1) = v_1|a_2) = y \neq x$

Essentially, Baseline Relevance says that you may judge a_1 to more valuable than a_2 by z and still regard a_1 as evidence for its own objective value. Everything I say about Relevance applies to Baseline Relevance since ii) and iv) are analogous to the learnability premise. (There is nothing surprising here—Ahmed and Spencer on p. 1178 say that the rationale behind Baseline Relevance is the same as that behind Relevance.) So, my discussion of Relevance applies equally to Baseline Relevance.

[89] Ahmed and Spencer's full argument allows us to reject Expectationism—which I think we clearly should in light of the Allais Preferences—and adopt a weaker thesis about how subjective and objective values relate:

> **Dominance**: If options have objective values then: if an agent's option set is A and a uniquely maximises objective value in A, then a uniquely maximises subjective value. (He 'subjective value' is U, or what I have been calling instrumental utility.)

They show that Baseline Relevance (see previous footnote) is incompatible with Dominance and EDT. Again, everything I say about Relevance applies to Baseline Relevance and hence to their Dominance argument.

Ahmed and Spencer each have different responses. Ahmed (p. 1184) claims that we should reject the existence of objective value, while Spencer (p. 1184) claims that we should reject EDT.

I do not think, however, that Ahmed and Spencer have demonstrated that EDT is incompatible with the existence of objective (or actual) value. Recall their learnability premise: you should be able to get evidence about objective value from your acts. Why accept this? We started with the idea that actual (or objective) value is determined by what holds independently of what you can do. *This* strikes me as the non-negotiable feature of actual value, that it is the kind of value that holds in virtue of the way the world is. In particular, we want to contrast objective value (what best promotes your ends given how the world is) with subjective value (what you take to best promote your ends given your uncertainty about how the world actually is). Objective value is just the kind of thing that contrasts with subjective value—for Lewis this will be the utility of an act in conjunction with the actual causal dependency hypothesis, while for someone like Horgan it will be the utility of the act in conjunction with the correct backtracking dependency hypothesis.

Why then think that objective value must have some additional learnability feature? On the view of EDT I have sketched, we follow Horgan in taking states to be act-independent (i.e., $C(k|a) = a$ for all a). So, the actual value of an act is the value that the actual backtracking-dependency hypothesis assigns to that act. And there is no act such that performing that act changes your credence in a backtracking-dependency hypothesis, hence there is no act such that performing that act provides evidence for its own objective value. If we accept this characterisation of actual value, then the evidentialist has no motivation for the learnability premise of Ahmed and Spencer's argument.[90]

[90] Ahmed and Spencer (2020, p. 1173) provide a case that they claim supports the learnability premise. That case involves your (i) deciding whether to block Shaft A or Shaft B to save some miners, (ii) having been told previously which shaft the miners are in, but (iii) having forgotten, though nonetheless thinking that your past knowledge will subconsciously influence you to block the shaft that the miners are in. Here, they claim that anybody who denies the learnability premise denies that 'the objective value of blocking [some shaft] depends on where the miners are. They would have to hold that … the objective value of blocking shaft A at a world in which the miners are in shaft A is the same as the objective value of blocking shaft A at a world in which the miners are in shaft B. But that's absurd.' The reason it is absurd is that (p. 1173) 'the objective value of blocking shaft A *depends on where the miners actually are*' (emphasis mine).
I do not think that this case shows that the learnability premise holds, for a few reasons: (i) Cases of forgetting strike me as most susceptible to the 'Tickle Defence'—if I have forgotten some fact, then once I have conditionalised on my initial *inclination* to, say, unblock Shaft A over Shaft B, it is unclear that my subsequent decision has any further evidential bearing, (ii) The case involves *forgetting*. But forgetting is not rational by Bayesian lights—it violates conditionalization. Cases of forgetting do not motivate the learnability premise, since I take that premise to govern the credences of a *rational* agent, and (iii) If we take the backtracking intuition seriously, it is unclear what we should say about objective value in Ahmed and Spencer's case. Say that in the actual world the miners are in Shaft A and you block Shaft A. If you had done otherwise, that *might* have been because the miners were in a different shaft, which subconsciously influenced you to act differently. If we take the influence of the miner's location on your

Ahmed and Spencer's response, I take it, would be that it is part of the meaning of 'objective' value that it be learnable. Hence 'value given the correct backtracking-dependency hypothesis' cannot be a good definition of objective value. Perhaps the idea is that if the learnability premise did not hold, objective value would not be 'out there' in the right way? But surely, if we take the intuition behind EDT seriously, we have no reason to think that objective value is 'out there' in this strong sense. a is objectively more valuable than b if $a \rightarrow o$ and $b \rightarrow o'$ hold where $o > o'$. And since the conditional → already takes each act's evidential bearing into account, I am unsure why we should insist on the learnability premise. The intuition behind the learnability premise, I suspect, presupposes the causalist notion of independence: if the causalist standard is the right one, then $C(a \rightarrow o) \neq C(a \rightarrow o|a)$ for some acts, outcomes, and rational credence functions. (Simply note that since your acts can provide evidence about propositions outside your causal influence, acts can provide evidence for the truth of non-backtracking or causal counterfactuals.) But if we deny the causalist standard, I am unsure how to motivate the learnability premise.[91]

Let us suppose, however, that the learnability thesis does hold. What dialectical force does this have? Ahmed claims that it gives us reason to reject the existence of objective value, while Spencer claims that it gives us reason to reject EDT (p. 1184): 'One of us is inclined to take the inconsistency to amount to a metaethical refutation of EDT. The other is inclined to take the inconsistency to amount to a metaethical refutation of the claim that options have objective values. Obviously, we cannot settle here whether to reject EDT or the claim that options have objective value. But one of them has got to go.'

But the argument above shows that this does not really get us very far. If we reject the existence of objective value, we can still define objective *schmalue* as the utility of o such that the backtracking counterfactual $a \rightarrow o$ holds in the actual world. Objective schmalue is distinct from subjective value (or instrumental utility), since it characterises the utility of acts not from your perspective of uncertainty but utility given how the world is. So objective schmalue does the

decision seriously, then the objective value of each act may reflect the past conditions that give rise to that act. Of course, at each *world* the utility of that world depends on where the miners are in that world (so, we might say that the value of the actual world is grounded in facts about where the miners are). But that is not to say that the objective value of each *option* settled only by where the miners are in the actual world—that ignores that your doing otherwise might involve the subconscious influence of the miners being elsewhere. So, in cases where it seems most obvious that actual value and 'value given the correct backtracking dependency hypothesis' diverge, it seems that the agent is either failing to conditionalise on relevant information *or* forgetting. If not, then that it is reasonable to take 'objective value' to be value under the correct backtracking counterfactual.

[91] Note that we can still learn about objective value on the view of EDT I have sketched, just not in the strong way the learnability premise requires. For example, by learning facts about the predictor's efficacy, you might learn about the objective value of one-boxing. Hence, objective value is still out there in the sense of being a mind-independent part of the world you can learn about. What I deny is that objective value must be the kind of thing you can learn about *merely by performing some act*.

work I want objective value for (though it does not obey the learnability premise). The evidentialist standard of rationality is therefore compatible with almost everything we want from a theory of objective value. So, contra Spencer, I cannot see that EDT's inconsistency with the existence of objective value would provide any reason to reject EDT—at least we would need some compelling argument that objective schmalue does not do the metaethical work we want. And, contra Ahmed, I do not think it says much that the EDT'er must reject the existence of objective value. At the very least they can accept the existence of objective schmalue—but given how close schmalue is to value, I suspect this just amounts to re-characterising objective value by rejecting the learnability premise. So, the argument from objective value does not break the deadlock. Again, it presupposes a notion of independence that the evidentialist already rejects.

2.3 Against Causalism I: Why Ain'cha Rich?

I have argued that EDT is not as easily refuted as some think.[92] We cannot simply appeal to a fully informed bystander or the actual contents of the opaque box to underwrite two-boxing. I now consider two well-known arguments against two-boxing. Again, the causalist can resist these arguments because each presupposes an understanding of independence that the causalist already rejects.[93]

The first argument is often expressed as a taunt: if you're so smart, why ain'cha rich? In all probability, the two-boxer walks away from Newcomb's Paradox poor, while the one-boxer walks away rich. And since the goal of acting rationality is to promote your ends, one-boxing does a better job of promoting your ends than two-boxing. This is the 'Why Ain'cha Rich?' (WAR) argument. Put slightly differently, you have no reason to follow a standard of rationality that leaves you poor.

[92] I have set aside a recent argument against EDT from Wells (2019), for which see Ahmed (2020) for a response.

[93] I will set aside various arguments leveraging dynamic exploitability, for example variants on the Money Pump in that exploit the way that U_C changes as the causalist performs (or considers performing) various acts (I recently learned of Gallow Forthcoming, which contains a very helpful discussion of such cases—see especially his discussion and the references in Section 3). This is partly due to a scepticism about the significance of such arguments, for which see Hedden (2015). Moreover, on the most sophisticated recent defence of Money Pumps due to Cantwell (2003), it is not clear that dynamic exploitability really does show that the causalist is irrational. Roughly: Money Pumps, if anything, show that your preferences are bad by your own lights (Cantwell 2003, p. 383). So, in cases where you are exploited due to preference change (even predictable preference change), it is not *your* preferences that are bad, but the way that your future preferences interact with your current choices that leads to exploitation. In standard cases where the causalist is exploitable, it is because your their provide evidence for their own efficacy, so their preferences change over the course of a sequence of decisions. So, even if the CDT'er suffers from dynamic exploitability, this is because of they way their preferences change, not because their preferences are incoherent.

But again, the devil is in the details. It is not, according to the causalist, literally true that *you* would be better off as a one-boxer. That claim is only true if we think about what would happen were you to one-box in a backtracking way. So, in what sense does the one-boxer do better than the two-boxer?

One thought is that one-boxers do better in general than two-boxers. But we must be careful here. If we take a one-boxer off the street and compare her with a two-boxer off the street, we are likely comparing agents with different causal histories. If the one-boxer is rich and the two-boxer is poor, this is because the one-boxer faced a predictor who placed a million in the opaque box, while the two-boxer faced a predictor who placed nothing in the opaque box. This means that the (rich) one- and (poor) two-boxer differ with respect to facts that hold independently of what they do. And as Bales (2018b) argues, it is unfair to criticise an agent for not doing as well as someone in a different situation. More precisely, Bales (2018b, p. 263) endorses:

> **Fair**: A comparison of the return to agents reflects the rationality of the decisions made by these agents (and hence is a fair comparison) only if the circumstances of all agents are the same with regards to everything except: (a) the agent's decision; and (b) those things that the agent's decision makes a difference to.

And, as Bales rightly notes, the causalist already thinks that the rich one-boxer and poor two-boxer differ with respect to facts that their decisions do not make a difference to. So, we cannot make a fair comparison and criticise the two-boxer for not doing as well as the one-boxer. In a similar spirit, Joyce (1999, p. 153) considers Rachel, a poor two-boxer, responding to the WAR taunt of Irene, a rich one-boxer: 'I'm just not like you Irene. Given that I know that I am the type who takes the money, and given that [the predictor] knows that I am this type, it was reasonable of me to think that the $1,000,000 was not in my account. The $1,000 was the most I was going to get no matter what I did. So the only reasonable thing for me to do was take it'. The two-boxer is not like the one-boxer—they are in a different situation because they face a different predictor. The two-boxer therefore does as well as they can, and it is unfair to criticise them for failing to do as well as the one-boxer.

A recent version of WAR, again due to Ahmed, dispenses with inter-agent comparisons and so survives criticisms from Bales and Joyce. It is worth quoting at length Ahmed, who also considers the one-boxer Irene comparing herself to a two-boxer Rachel:

> [N]either Irene nor any other agent cares, when choosing, about whether she does *better* than anyone, actual or possible, counterpart or not. Nor does she care about whether she will consequently regret her choice. She only cares about her terminal wealth. So Rachel's

> point is irrelevant [i.e., the point that Rachel, as a two-boxer, did as well as she could holding fixed the predictor's prediction]. … What matters is only that she [Irene, the one-boxer] is *richer*. (Ahmed 2014b, pp. 185-6)

Note that Ahmed accepts that the two-boxer did as well as they could have in some sense—their one-boxing counterpart is worse off than they are. But he denies that what matters is how well you do compared to some other agent—who cares about what could or would have happened? Instead, what matters is that the one-boxer is foreseeably richer than the two-boxer (where this is compatible with the claim that the one-boxer would have done better by two-boxing and the two-boxer would have done worse by one-boxing).

But what does it mean that one-boxing 'foreseeably does better'? Ahmed has ruled out its being a comparative notion between counterparts, so what is it? One thought is just that the probability of being rich supposing that you one-box is higher than the probability of being rich supposing that you two-box. But this presupposes that we know what the relevant form of 'supposition' is, and Joyce (1999, Chapter 7) demonstrates that there are forms of supposition on which the claim is false. Elsewhere, Ahmed (2014b, p. 181; see also Ahmed and Price 2012) formulates the WAR as:

1. The average return to one-boxing exceeds that of two-boxing. (*premise*)
2. Everyone can see that (1) is true. (*premise*)
3. Therefore one-boxing foreseeably does better than two-boxing. (*subconclusion*)
4. Therefore CDT is committed to the foreseeably worse option for anyone facing Newcomb's problem (*by 3*). (*conclusion*)

Crucially, this argument makes no inter-agent comparisons. So, the Joyce-Bales point—that rich one-boxers and poor two-boxers differ with respect to facts outside their control—has no bite. Doing foreseeably better is just about having a *higher average return*.

But again, the causalist can resist the 'average return' version of WAR, this time by rejecting Premise 1 (and thereby Premise 2).[94] We need to get clear on what the 'average return' of an act is. Ahmed and Price (2012, p. 17) define it as the 'average returns (AR) … over many trials', and they assume that $AR(a) = \sum_o Cr(o|a) \cdot u(o)$. There are two assumptions here: i) that an act's

[94] We might also question Premise 3. After all, a risk-averse act has a lower average return than a risk-neutral one, but many think that there is a reasonable sense in which the risk-averse act does foreseeably better. The relevance of repeated trials to a one-shot decision has been discussed elsewhere (see, for example, Baron 2009 (p. 243), Briggs 2019, and Hájek Forthcoming[a] for criticisms of the relevance of long-run arguments). I grant Premise 3 for the sake of argument—even *if* long run arguments are relevant, it is not clear that they support the one-boxing verdict in Newcomb's Paradox.

average return is its average return over many trials, and ii) this is equal to its probability-weighted utility, where probabilities are conditionalised on the act's performance.

Granting the first assumption as a stipulative matter, why accept the second? If we imagine many trials of the decision you find yourself in Newcomb's Paradox, then the average return of two-boxing exceeds that of one-boxing by a thousand dollars. After all, every time you repeat the decision you currently face, you get a thousand more by two-boxing. So, if we think about repeated trials of *this* decision, it is reasonable to hold the contents of the opaque box fixed—the contents of the box are settled as a matter of historical fact and not as some stochastic process that is part of the decision you face. So, if we hold fixed the contents of the box over the many trials, it is not true that the average return of one-boxing exceeds that of two-boxing, hence Premise 1 is false. Ahmed and Price must therefore intend Premise 1 to be taken in some other sense (and, for Premise 2 to hold, everyone must be able to see that this is the right reading!).[95] Perhaps the average return of two-boxing exceeds that of one-boxing in decisions relevantly similar to yours?

But this only pushes the bump under the rug, since we have to say what relevant similarity is. An attractive idea is that a decision is relevantly similar to yours if it is the same with respect to facts that hold independently of your decision. But again, this mentions independence, which the causalist thinks means causal independence. And if we hold fixed facts causally independent of yours, then we hold fixed the predictor's decision and therefore the contents of the opaque box (even if the predictor is an accurate predictor of *my* decisions, they have made their choice, and changing my decision now won't affect the basis on which they made their decision). This means the average return of two-boxing is again a thousand dollars higher than that of one-boxing. So, again since the causalist reasonably rejects the evidentialist's standard for independence, they can maintain that the average return of two-boxing is higher. The causalist chalks the appeal of this version of WAR up to a reference class fallacy: in a wide reference class containing typical one- and two-boxers, the one-boxers do better; but in a narrow reference class containing agents with the same causal history as you, the two-boxers do better. So, the causalist can deny that two-boxers do worse than one-boxers in any relevant sense.

[95] Ahmed (2014b, p. 181) argues that the expected return of one-boxing is higher than two-boxing, though of course this claim presupposes that 'expected return' is not explicated in terms of U_C. Ahmed (p. 181) thinks that we should calculate expected return using relative frequencies—if one-boxing makes you a millionaire nine times out of ten, the expected return of one-boxing is $.9 \cdot u(\text{Million})$. But this will not persuade the causalist. While their conditional credences capture facts about the frequency of the predictor's success, they deny that expectations calculated based on these relative frequencies directly inform what *you* should do in *this* decision situation. Relative frequencies encode facts about the predictor's long-run performance, but what matters in Newcomb's Paradox is what is best for you in *this* decision situation.

Perhaps the evidentialist can push back. They might claim that we should allow the predictor's decision to vary across trials since they are an excellent predictor—typical one-boxers face a benevolent predictor, and so we should assess one-boxing relative to the *typical* past conditions that you face if you one-box. Perhaps then what matters are the typical conditions that give rise to each act?

That claim by itself is surely false. Say that I know that the roulette wheel is rigged in my favour—the casino manager is usually nasty to me, but on this occasion she is being atypically nice. Thinking then about playing roulette relative to typical conditions (even those that lead me to play roulette) makes the average return of playing worse than that of not playing. But if I know the wheel is rigged in my favour, I should play—we should hold fixed the atypical behaviour of the casino manager! So, thinking about the average return of an act cannot be a matter of re-running the same decision repeatedly and allowing past facts that are independent of my choice to vary. What really matters is that the manager's behaviour is fixed independent of my decision (this is true on the causalist and evidentialist standard), so we should hold it fixed when deliberating. So, what matters is not how the typical one- or two-boxer does, but how such agents do holding fixed what holds independently of their decision. And the causalist thinks that past facts (e.g., facts about the predictor's decision) hold independently of what you do, so it is impermissible to assess the average return of one- and two-boxing by 'fitting the past' to each act. Anyone attracted to causalism will say that you should hold the past fixed when assessing the average return of one- and two-boxing. And when we do so, the causalist does better.

But this may strike you all as a dodge. Perhaps Ahmed and Price's Premise 2 kicks in: whatever senses 'the average' might have, we can *just see* that the average return of one-boxing is higher. I am not so sure. Ahmed and Price's WAR asks us to think about repeated trials of an act. But it is not at all clear to me how to think about Newcomb's Paradox, which is a one-shot decision, in the context of a repeated trial. If I imagine repeated trials of Newcomb's Paradox, I may think that what I do at a given choice will influence the predictor's decision at future choices; in particular, by one-boxing now, I may make the predictor more likely to predict one-boxing in the future. Or, I can imagine a repeated trial of Newcomb cases exactly like mine, meaning that the predictor makes their prediction in each trial relative to my current track record and past facts about my psychology (and not the track record and psychology of some future self who has made more decisions than me, some of which may be to one-box). Given these two ways of thinking about repeated trials, the causalist can reasonably insist that the latter is the right way of thinking, given the one-shot nature of Newcomb's Paradox—I'll *never* see the predictor again, and this is my *only* chance to get the money, so I should think about a repeated trial facing *this* predictor and intervene to get the extra thousand in *these* circumstances.[96] The initial question of what to hold fixed when deliberating infects even sophisticated versions of WAR.

2.4 Against Causalism II: Instability

Another classic argument against CDT is the argument from *instability*. CDT can tell you to do something that it judges to be irrational on performance. Some reject CDT because of this. To illustrate, consider the following case from Gibbard and Harper (1978):

> *Death in Damascus*: You must go to either Aleppo or Damascus. Death is a cunning and highly accurate predictor of your choices (Death has previously guessed ninety percent of your choices correctly). Death went to either Aleppo or Damascus this morning. If you avoid Death you live; if you meet Death you die.

We can model this:

	Death in Aleppo	Death in Damascus
Go to Aleppo (A)	0	10
Go to Damascus (D)	10	0

Say that initially:

$$C(\text{Death in Aleppo}) = C(\text{Death in Damascus}) = .5$$

With conditional credences:

$$C(\text{Death in Aleppo}|\text{A}) = C(\text{Death in Damascus}|\text{D}) = .9$$

Now consider CDT's advice:

$$U_C(\text{Aleppo}) = U_C(\text{Damascus}) = 5$$

[96] Ahmed (2014b, p. 182) illustrates the WAR argument by imagining a 'one-boxer explaining patiently to the poor causalist just what he needs to do to become a millionaire … one-box. And we can imagine the causalist repeatedly doing just the opposite, with unfortunate results.' But we must be careful—if two-boxers do worse than one-boxers *over the course of multiple, distinct decision situations*, this may not be an argument against *causalism*. Rather, it is to assess individual acts (e.g., two-box) as parts of *plans* or *policies*, which involve repeated acts. But (i) whether you should causally promote the good, and (ii) whether you should choose based of which *act* or *policy* promotes the good, are separate questions. Meacham (2010) discusses a causal view that evaluates not acts but strategies—I discuss this in Chapter 4. What matters here is that long-run considerations illustrate the effects of policies, but my focus in this chapter is on individual acts.

CDT, reasonably enough, says that given your symmetric credences and the symmetric payoffs in the case, you may go to either city. But note what happens when you decide, say, to go to Aleppo. Since $C(\text{Death in Aleppo}|\text{Aleppo}) = .9$, your updated credences mean that you now evaluate:

$$U_C(\text{Aleppo}) = 1 < U_C(\text{Damascus}) = 9$$

So, CDT permits you to go to Aleppo, but *on following CDT's own advice* it tells you not to go to Aleppo. Odd indeed! Surely doing something rational cannot render that very thing irrational. Yet this is what CDT says, and some therefore reject CDT.[97]

But the causalist has a simple response. Firstly, they can note that there is nothing incoherent in CDT's advice—claiming that something is rational now does not commit you to thinking that it will be rational in the future. If we receive new information or our tastes change, then what is rational can change. What does seem wrong is if you have stable tastes and a decision theory changes its advice in the absence of new evidence. In that case, it really must be that the theory's pre- or post-choice advice is mistaken. Fortunately, however, Death in Damascus is a case in which you *do* receive new information (moreover information that is relevant to the outcome of your decision), namely evidence about where Death is. So, CDT is guilty of nothing more than changing its advice in response to new evidence. Gibbard and Harper (1978) accept that CDT's verdict in Death in Damascus odd, but they claim that this is because Death in Damascus itself is odd. When playing against a cunning predictor like Death, we have no good reason to think that the correct decision theory will offer the same advice pre- and post-choice (cf. Hare and Hedden 2012, Section 3).

But you might maintain that there something odd about a decision theory changing its advice in response to evidence that you could have seen coming. In Death in Damascus, CDT does not change its advice because of some 'external shock'—a news report or an eyewitness who saw Death. Rather, it changes its advice in response to evidence that you knew you would receive all along. Surely CDT should not change its verdict in light of this predictable kind of evidence?[98]

That objection, however, smuggles in a form of backtracking reasoning. In Death in Damascus, you are undecided where to go. In particular, for your credences to be coherent, it must be that $C(A) = C(D) = .5$. So, when CDT says that you may go to Aleppo, this reflects the fact that

[97] Richter (1984) is a classic early statement of the argument. Weirich (1985), Harper (1986), Sobel (1988c) each amend CDT to say that in cases where no act is stable there is no rational act. Spencer and Wells (2019) reject CDT but adopt a view that recommends two-boxing.

[98] Thanks to Melissa Fusco for discussion here.

you do not (yet) think you will go to Aleppo. It is therefore false that you currently believe that you will receive evidence that Death is in Aleppo. Given that you are currently unsure of your going to Aleppo, it makes sense to recommend that you go to Aleppo—do the unexpected! Of course, you could evaluate Aleppo relative to the news that going to Aleppo would provide. But that would be to evaluate Aleppo not by your current lights but by the lights you have if go to Aleppo. And that would be backtracking reasoning—simply note that the causalist only assigns $.5$ credence to the conditional 'If I go to Aleppo, Death is in Aleppo'. The causalist attaches no significance to the evidential bearing of A.

Even if instability is not itself incoherent, you might think that there is something strange in the fact that whatever you do, CDT will tell you that you have done the wrong thing. Richter expresses this concern:

> [T]he causal theorist isn't merely making the innocuous claim that [you] have no rational alternative; [they are] committed to the claim that whatever [you] do in this case is irrational, i.e., not permitted by rationality. Whatever we mean by 'rational' we surely don't want to allow that an agent could be in a situation in which every available alternative would be irrational. (Richter 1984, p. 395)

But the discussion so far shows that Richter is equivocating. There is no time at which CDT says that all acts are irrational; at each time CDT ranks acts by causal expected utility, so at each time there is a maximal option which CDT judges to be rational. So whatever credences you have, CDT judges at least one available alternative to be rational.

Perhaps Richter is taking issue with the fact that whatever CDT recommends, it *will* tell you that you are irrational to do that thing (when you do it). For example, if you decide to go to Aleppo (Damascus), then you receive evidence that Death is in Aleppo (Damascus)—so whatever you decide, CDT tells you that you are irrational on implementing that decision. So, CDT gives rise to a kind of rational dilemma: for any option you might adopt, there is some time at which you *will* judge that option to be irrational. And the correct decision theory should never give rise to a rational dilemma.

In response, the causalist can deny that this is a normatively significant dilemma. Let's say we take the view that options are decisions, so what counts as an option supervenes on your present mental states (see Weirich 1983 and Hedden 2012). If so, then while CDT might tell you to do something the *implementation* of which you judge to be irrational (after having made that decision), the decision itself is rational. And it is the decision itself, not the later implementation, that counts as an option for you. After all, 'deciding to ϕ' and 'carrying out ϕ at some future time' are distinct. So, that CDT will judge some future act (even one that follows from your current decision) irrational does not constitute a rational dilemma for *you*—decision theory applies to options, so it is irrelevant what CDT will say about future events that are not options.[99]

Spencer and Wells (2019) argue against CDT along similar lines. They claim (2019, p. 38) claim that since the output of decision theory is supposed to be action guiding, if a decision theory tells you to maximise some *value quantity* (such as U_C), then you must have *stable access* to that value quantity. So, if 'Maximise U_C' is the correct decision rule, agents must always be in a position to (i) know that some option maximises U_C, and (ii) know that that option maximises U_C conditional on its performance. They argue for (ii) on the basis that if some agent is not in a position to know that an option maximises U_C conditional on performance, she is either 'surprised by which option she chooses … or anticipates choosing an option that is not [U_C] maximising, in which case her choice is not guided by U_C' (p. 38).

But this line of argument rests on the same equivocation as above. We ought not infer from 'thou shalt maximise U_C' to the stronger 'thou shalt do something that each of your time-slices takes to maximise U_C'. You are still being guided by U_C-maximisation if you do something *now* that binds you to doing something that, by your post-choice lights, fails to maximise U_C. If you *now* take something to maximise U_C but will later think it fails to maximise U_C, then so much the worse for your later self (whose views you think are likely mistaken). To be sure, instability does raise problems for how to *implement* CDT's advice, since our time-slices are in competition (I return to this point in Chapter 4). But at least in some cases, I assume we can commit to doing what CDT recommends at a time, perhaps by forming an intention to do that thing (see Bales 2020), or simply because some acts are things we can directly implement without giving our future selves opportunity for reconsideration. And in those cases, I see no problem with following U_C-maximisation, though you lack stable access in Spencer and Wells' sense.[100]

[99] Sobel (1988d, p. 199) claims that when there are no stable acts, no act is rational *or* irrational. This is because you cannot 'make up your mind' to perform an unstable act—Sobel (p. 199) thinks that "possession of [credences] and preferences that are settled and constant for a period leading up to a time is a necessary condition to any of an agent's actions' at this time being either rational or irrational". But again this seems to rest on an equivocation. If we think of your options as decisions, then we have no reason to think that a rational option must have been rational (by your past lights) shortly before performance, nor must it be rational shortly after performance—what matters is that *each time* there is something that you take to be permissible at that time.

[100] This brings us to a background disagreement. Spencer and Wells think that if you rationally ought to do something, you are in a position to *know* that it maximises U_C (or whichever quantity you take to be action-guiding). I am not convinced that knowledge need enter the picture here—I am happy to say that you ought to do the things that you currently take to be the best means to your ends, even if you do not *know* that they maximise, say, U_C in Spencer and Wells' sense.

To be clear, in Chapter 4 I will return to instability and argue that there are indeed pragmatic problems in knowing how to *interpret* CDT's advice in cases of instability. I will argue that instability provides *pragmatic* reasons to adopt EDT over CDT for action-guiding purposes, though this is compatible with CDT representing a legitimate standard of rationality. What matters at this stage is that unstable advice is not incoherent—nothing about CDT's advice demonstrates in a non-circular way that CDT's advice is irrational. The causalist can reasonably accept instability as a feature, not a bug, of their view—indeed, they can maintain that stable advice is the *wrong* kind of advice to expect when playing against Death.

2.4.1 Briggs' Self-Sovereignty

There have been a range of rehabilitations of the argument from instability. It would be impossible to assess them all. But I will outline two prominent recent rehabilitations, Briggs' *Self-Sovereignty Argument* (2010b) and Egan's (2007) argument from cases of *Extreme Instability*, and show that both can be dealt with in much the same manner as the standard argument.

Briggs (2010b) argues against CDT based on a *Self-Sovereignty* principle. Briggs' key idea is to think of your possible future selves as voters and then to formulate principles governing your current choices based on those future selves' preferences. Let $U^a(\cdot)$ be a measure of instrumental utility by the lights of your future self who does a. For example, Briggs defines (p. 19, notation adjusted):

$$U_C^a(b) = \sum_k C(k|a) \cdot u(o_{b,k})$$

That is, your future a-self's causal expected utility for b is the probability-weighted average of each outcome, where the probability-weights include the fact that you have learned that you do a. Similarly, Briggs suggests defining (p. 19, notation adjusted):

$$U_E^a(b) = \sum_k C(k|a\&b) \cdot u(o_{b,k})$$ [101]

[101] Briggs acknowledges that this is not an unproblematic quantity, since we conditionalise on $a\&b$, though a and b are distinct acts, hence incompatible. We could either (i) adopt Popper functions, or (ii) adopt a 'trembling hands' approach and conditionalise on $a^*\&b$, where a^* is a decision to do a that allows for some chance of failure and doing b instead.

Again, your future a-self's evidential expected utility for act b is the probability weighted average of each outcome, where the probability weights reflect your credences after having performed (or deciding to have performed) a. Briggs (p. 22) introduces two conditions and argues that both should hold on any reasonable measure of instrumental utility, U:

> **Pareto:** If for all options b, a_1, a_2, $U^b(a_1) \geq U^b(a_2)$, then $U(a_1) \geq U(a_2)$. Moreover, if for some option b, $U^b(a_1) > U^b(a_2)$, then $U(a_1) > U(a_2)$.
>
> **Self-Sovereignty:** Take two agents with credence functions C and C' and instrumental utility functions U and U', respectively, facing a decision situation with the same dependency hypotheses and possible acts. For any two acts a_1, a_2, if $U^{a_1}(a_1) = U'^{a_1}(a_1)$ and $U^{a_2}(a_2) = U'^{a_2}(a_2)$, then $U(a_1) \geq U(a_2)$ if and only if $U'(a_1) \geq U'(a_2)$.[102]

Briggs argues that the first principle rules out EDT, since each possible future self prefers you to two-box. (Though the previous discussion shows that, despite appearances, a lot is smuggled into Pareto: it requires that you hold fixed what each future self holds fixed, including the contents of the opaque box. But the evidentialist takes this to ignore relevant statistical dependence relations.)

What matters here is that Self-Sovereignty rules out CDT because CDT gives unstable advice. To see this, note that Self-Sovereignty is a kind of deference constraint. If two agents have different credences, but their future a_1-selves agree about the goodness of a_1 and their future a_2-selves agree about the goodness of a_2, then those two agents should agree about how to compare a_1 and a_2. Present disagreements are swept away by future agreements about the utility of acts. Note then that in Death in Damascus $U_C^{\text{Aleppo}}(\text{Aleppo}) = 1 = U_C^{\text{Damascus}}(\text{Damascus}) = 1$. Given that your current causal expected utility function is U_C, let $U_C{}'$ be the causal expected utility function of some possible future self who is inclined to go to Aleppo with credence function C'. Since $C'(\text{Aleppo}) > C'(\text{Damascus})$, we get $U_C'(\text{Aleppo}) < U_C'(\text{Damascus})$. Provided that your future self has the same conditional credences as you, we know $U'^{\text{Aleppo}}(\text{Aleppo}) = 1 = U'^{\text{Damascus}}(\text{Damascus}) = 1$. Therefore $U^{\text{Aleppo}}(\text{Aleppo}) = U'^{\text{Aleppo}}(\text{Aleppo})$ and $U^{\text{Damascus}}(\text{Damascus}) = U'^{\text{Damascus}}(\text{Damascus})$, so by Self-

[102] This is actually slightly weaker than Briggs' condition. Briggs formulates the condition to allow for agents to have different credences *and* face different dependency hypotheses. The constraint in the main text is sufficient for present purposes, and everything I say here applies equally to Briggs' stronger constraint.

Sovereignty *you* currently prefer Aleppo to Damascus if and only if your future self (who inclines towards Aleppo) prefers Aleppo to Damascus. But CDT requires that you be indifferent between the two, while your future self (who inclines towards Aleppo) strictly prefers Damascus. So, CDT violates Self-Sovereignty.

In short, Self-Sovereignty says that your a_1-self is the authority with respect to the goodness of a_1, as your a_2-self is the authority with respect to the goodness of a_2. But CDT rejects this thought and says that your current views trump the views of your future selves.

Why accept Self-Sovereignty? Briggs (p. 23) points to the intuitive idea that you make decisions 'to end up with a future self who is in some sense happy to exist', and this means giving veto power to each future self—if doing a is bad by the lights of your a-self, then you have decisive reason not to do a. (Analogy: there are lots of facts that might count in favour of buying my friend Aden a beer, but if I am acting to make him better off and *he* tells me he does not want a beer, then I have decisive reason not to buy the beer.)

But the causalist will not be convinced by this argument. Perhaps we act to make our future selves better off, but that does not require deferring to our future selves in every respect. Perhaps in matters of *taste* I should defer to my future self (though even this might be contested, cf. Hare and Hedden 2012). But crucially, CDT's advice is unstable because you anticipate changes in your *epistemic* situation. And acting in somebody's interests does not entail deferring to their credences. (Analogy: if I discover the reason that Aden does not want the beer is that he falsely believes he must drive home, then I might still buy him the beer.) The question becomes whether to defer to your future selves, and indeed which future selves to defer to when you and those selves have different bodies of evidence.

Self-Sovereignty says that you should defer to your future a-self about the choiceworthiness of -a. But this is *prima facie* odd. Recall that going to Aleppo provides strong evidence that Aleppo is bad—$U^{\text{Aleppo}}(\text{Aleppo}) = 1$. You judge that this utility assignment is accurate only if you actually go to Aleppo (if you do not, then you have no reason to take the evidence that going to Aleppo would provide seriously). But as mentioned before, you do not think yourself likely to go to Aleppo since you could just as well go to Damascus. Indeed, in Death in Damascus, there is a subtle form of incoherence in treating your future Aleppo-self as an expert with respect to Aleppo and your future Damascus-self with respect to Damascus. Since you cannot perform *both* acts, your Aleppo-self's views are accurate if and only if your Damascus-self's views are not. So, to defer to your Aleppo-self with respect to the instrumental utility of Aleppo is to treat your

Damascus-self as untrustworthy; and to defer to your Damascus-self with respect to the instrumental utility of Damascus is to treat your Aleppo-self as untrustworthy. How then can you treat your Aleppo-self as an expert (with respect to the utility of Aleppo) and your Damascus-self as an expert with (with respect to the utility of Damascus)? You can only do so if you assess each act relative to different credence functions, reflecting different views about how the past is—and this is precisely what the causalist (reasonably) denies you should do. Relative to *your* current views, you cannot simultaneously treat your Aleppo-self and Damascus-self as experts.

The causalist says that the correct thing to do is to weight each possible future self's advice by how accurate you expect them to be, which is determined by how likely you are to perform that act.[103] Every future self gets a vote. But you recognise that your various future selves have different views about how the world is, and you need to balance those competing, conflicting pieces of advice. Since you are currently undecided about where to go, neither your credence function conditionalised on going to Aleppo nor the one conditionalised on going to Damascus is better than your current unconditional credence function.

This brings us to the crux of my disagreement with Briggs. They claim, in line with the voting metaphor, that rational agents should ignore 'how likely [each future self] is to come into existence by the agent's current lights'. One person, one vote, no weighting. But the causalist rejects this principle. I care about how likely each future self is to come into existence simply because I care about whether they are reliable or not. If I am likely to go to Aleppo, then my Damascus-self's views are likely mistaken, and I have every reason to ignore them (even with respect to the instrumental utility of going to Damascus).

Briggs (p. 18) calls this line of thinking 'paternalistic'. Perhaps, but it is not a worrying form of paternalism. Rabinowicz provides a good motivation for overriding the views of your future self:

> '[You deem] it highly probable that these new beliefs, while rational, would be false. Therefore, [you] have no good reason to take such most probably mistaken hypothetical beliefs of [yours] into consideration.' (Rabinowicz 1985, p. 189)

[103] Formally:

$$\begin{aligned} U_C(a) &= \sum_k C(k) \cdot u(o_{a,k}) \\ &= \sum_{k,i} C(a_i) C(k|a_i) \cdot u\big(o_{a,k}\big) \\ &= \sum_i C(a_i) \cdot U_C^{a_i}(a) \end{aligned}$$

So, CDT's advice is a weighted average your possible future views, weighted by how likely each future self is to come into existence.

The intuition behind Briggs' charge of paternalism, I suspect, is that *if* you go to Damascus, then you get strong evidence that Death is in Damascus and your Damascus-self will think you should pay attention to this evidence. But again, the causalist simply notes that this smuggles in the wrong kind of conditional reasoning: they deny (the rational relevance of) the claim, 'if I go to Damascus, my Damascus-self's views are accurate', just as they deny (the rational relevance of) the claim that 'if I go to Damascus, Death is likely in Damascus'. Endorsing those claims would be to assess each act given your views on how the world likely is if you perform that act—*verboten*! The causalist should therefore embrace paternalism (in Death in Damascus). And insofar as the voting-paradigm suggests equal weighting for each future-self's vote, the causalist will reject the voting-paradigm.

I should note that Briggs is motivated not merely by ordinary cases of instability. They discuss a case due to Egan (2007) in which many think CDT goes wrong not simply by recommending an unstable act but a particular *bad* unstable act. I now turn to that case.

2.4.2 Egan's Psychopath Button

Consider the following (from Egan 2007, p. 97):[104]

> *The Psychopath Button*: In front of you is a button that, if pressed, kills all psychopaths. You would like to live in a world without psychopaths. Before pushing the button, however, a thought occurs to you: in all probability only a psychopath would push the button! You are certain that pushing the button does not make you a psychopath, just that pushing the button is the kind of thing that, in all probability, only a psychopath would do. You prefer living in a world with psychopaths to dying. What should you do?

We can represent this case as:

	Psychopath	Not Psychopath
Push	x	y
Refrain	0	0

With the conditional credence:

$$C(\text{Psychopath}|\text{Push}) = .8$$

[104] Note that a structurally similar case is discussed by Richter (1984, p. 400).

The problem specifies that $y > 0 > x$.[105] We then get that $U_C(\text{Push}) > U_C(\text{Refrain})$ whenever:

$$C(\text{Psychopath}) < -\frac{y}{x-y}$$

This means that you should Push whenever you are reasonably confident that you are not a psychopath. For example, following Joyce's (2012) discussion:

	Psychopath	Not Psychopath
Push	-30	10
Refrain	0	0

Here you should Push whenever $C(\text{Psychopath}) < \frac{1}{4}$.

Many, even those sympathetic to CDT, think that this is foolish. Pushing the button leads to your near-certain death, and following the correct decision theory should not lead to your near-certain death (not when there is an available alternative). CDT therefore cannot be the correct decision theory.

Some reject CDT because of Egan's case, including Egan (2007) himself. Gustafsson (2011) and Wedgewood (2013) provide updated decision theories that sometimes side with CDT and sometimes side with EDT, both of which recommend that you Refrain in the Psychopath Button. I will argue that Egan's case does not break the stalemate between causalist and evidentialist—the CDT'er need not modify their position to handle Egan-cases.

The first question to ask is what the Psychopath Button adds to ordinary cases of instability. The key feature is that it is highly asymmetric. Not only is CDT's advice unstable, it recommends an option that appears to be far worse than an available alternative (contrast this with Death in Damascus where both options are bad, but neither stands out as worse than the other). Here is how Egan makes this point:

> In general, when you are faced with a choice of two options, it's irrational to choose the one that you confidently expect will cause the worse outcome. Causal decision theory endorses … pressing. In general, causal decision theory endorses, in these kinds of cases, an irrational policy of performing the action which one confidently expects will cause the worse outcome. The correct theory of rational decision will not endorse irrational actions or policies. So causal decision theory is not the correct theory of rational decision. (Egan 2007, pp. 97-98)

[105] Setting the utility of the status quo to 0 simplifies calculations, but nothing substantive hinges on this assumption.

Egan's argument is that:

1. If you confidently expect that an act will cause a worse outcome (than an available alternative), then you should not perform that act.
2. You confidently expect that pushing will cause a worse outcome than refraining.

So:

3. You should not push.

This seemingly captures the asymmetry in the Psychopath Button and diagnoses what is wrong with CDT's verdict.

But again the causalist can push back, in this case by rejecting Premise 2. Given their current credences (which reflect their current views about how each act causally promotes the good), the causalist does *not* believe themselves to be a psychopath. So, they expect the causal consequences of pushing to be *better* than refraining. So, for Egan to assert that '[CDT] endorses … an irrational policy of performing the action which one confidently *expects* will cause the worse outcome [emphasis mine]', he must not be working with 'expectations' in the same sense as the causalist.

What other notion of 'expectation' could Egan have in mind? The obvious possibility is that *given* your pushing, you expect Push to cause worse outcomes than Refrain. But this proves too much, for given your refraining, Refrain has worse expected outcomes than Push. That is just to say that both acts are unstable, which does not capture the asymmetry that drives Egan's intuitions. What Egan really must mean then is that pushing (given that you push) has worse expected consequences than refraining (given that you refrain).

But this is just evidentialist reasoning! It is to assess each act based on how you take the world to be *if* you perform that act. And, at the risk of sounding like a broken record, this is backtracking. So, the causalist should not buy into the supposed asymmetry between pushing and refraining. Both are unstable, but only from the evidentialist perspective is pushing worse than refraining.

You might, however, worry that this response misses the point. We should not expect the Psychopath Button to show that there is something wrong with pushing by the causalist's own lights. Rather, we might take the case to show that the causalist's lights are the wrong ones to use. Of course, the causalist can rationalise pushing the Psychopath Button. But whatever story they might tell—pushing is just like any other unstable act, your current credences say that it is best, and so on—pushing is *clearly* the wrong thing to do, and so we should reject the causalist story in favour of a better one.[106]

I think, however, that we should be more circumspect about our intuitions in the Psychopath Button.

2.4.2.1 Biases and Heuristics

There are a range of well-documented biases and heuristics that drive our decision-making. Take, for example, the affect heuristic. We tend to focus on emotionally charged features of a case, especially negative ones. One particularly relevant consequence of this is that we overestimate the chance of emotive outcomes.[107] Death, I assume, counts as a negative and emotionally charged outcome. Insofar as we feel dread towards death, a large body of literature suggests that we are particularly sensitive to it and likely to overweight its probability.

Or consider Kahneman and Tversky's (1979) *framing effects*. We frame outcomes relative to a psychological 'status quo' and evaluate decisions relative to that frame. We are then loss-averse with respect to that status quo—we place more weight on avoiding outcomes that fall below the status quo than seeking outcomes above the status quo. It is reasonable to assume that the status quo in Egan's case is survival (that is the way the world would—I hope—be, had you never faced the case). Since survival is a good, or at least not a negative, consequence, we might posit that status quo reasoning and framing effects push us away from taking risks in Egan's case. Eriksson and Rabinowicz (2013, p. 822) note also that we employ the maximin heuristic—we act so as to minimize the severity of the worst outcome. Again, there is an easy way of avoiding death in the Psychopath Button: don't push, and there is no risk to your life!

There are plenty of other biases and heuristics we might point to. All I want to emphasise at this stage is that there are several reasons why you might not want the push, and these reasons have nothing to do with whether CDT is the correct theory of instrumental rationality. In general, extremely asymmetric cases of instability present an easy way of avoiding a severe loss, and insofar as biases and heuristics explain why we latch onto these options, we can explain why CDT's verdict seems puzzling. We do, of course, need to be careful when making empirical

[106] This is suggested by Egan (2007, p. 98). See also Bales (2020, p. 794) and Briggs (2010b, p. 9).

[107] For a helpful summary see Slovic & Peters (2006). Fischoff et al. (1978) argue that *dread* is a major influence in our assessment of risk. For further discussion see especially Finucane et al. (2000), Slovic et al. (2002), and Keller et al. (2006).

claims about what is or is not driving a response to a particular case, but we must also avoid rushing to conclusions about what rationality requires when confounding factors are in play.[108]

2.4.2.2 Rational Risk-Aversion

The causalist can go further—they need not appeal to *ir*rational biases or heuristics to explain away the 'Don't Push!' intuition in the Psychopath Button. CDT as often presented in the literature has two commitments: a causal definition of states and EU-Maximisation. But I do not want to tether CDT to EU-Maximisation. So, even if the Psychopath Button is a counterexample to CDT as usually formulated, this might just be because EUT is false.

Consider the certainty effect: if we reduce the probability of a good outcome by a fixed amount (say, 2%), we tend to be more concerned with that reduction when it shifts us away from *certainty* of the outcome (i.e., 100% to 98%), compared to shifting us away from some lower probability of the outcome (say, 10% to 8%). If you are an EU Theorist, then you take the certainty effect to be a deviation from ideal rationality (something other than probability-weighted utilities matters to you). But non-EU theories may vindicate something close to the certainty effect. Take WLU, which I discussed in Chapter 1, with a globally risk-averse weight function. Because outcomes are weighted by probability *and* relative weight, a decreasing weight function is highly sensitive to increases in the probability of a bad outcome. In the special case that one act guarantees a reasonably good outcome (say, your survival), then introducing a $.2$ probability of an extremely bad outcome (say, death) may decrease the value of that option significantly, even when combined with a $.8$ probability of some reasonably good outcome. And there is nothing irrational about this. Given the story of rational risk-aversion I told in Chapter 1, prioritising a sure-thing of a moderately good outcome and refusing to take small chances on bad outcomes is a reasonable way of achieving your ends. So, we might say that the certainty effect 'explains' our intuitions in the Psychopath Button; but I would add that plausible theories of means-end rationality permit (something close to) the certainty effect.

Even the EU Theorist will accept that a limited kind of risk-aversion is rational—the kind characterised by a concave utility function with respect to goods. Recall that in Joyce's modelling

[108] Egan presents another case, the Murder Lesion, that is structurally similar to the Psychopath Button but involves (i) a less happy status quo (life under a cruel dictator), and (ii) a dreadful, but less so, worst case scenario (imprisonment). Egan notes anecdotally that people have less strong intuitions about the Murder Lesion. If so, then this is some evidence that the factors I have pointed to are playing a role in driving people's intuitions.

of the case, he employed a utility function with $u(\text{Death}) = -30$, $u(\text{Status Quo}) = 0$, and $u(\text{Kill Psychos}) = 10$. I do not think that my utility function has these characteristics. It says that in ordinary contexts I am prepared to risk my life for a one-in-four shot at killing all psychos. No way! For one, I have never really put killing psychopaths at the top of my priority list—I rarely think about them, don't understand psychopathy well, and am unsure whether they are a net social good or not. Then there are various moral qualms about executing thousands without trial, and so on. So, for the Psychopath Button to be a case that *I* can make a judgement on, I would need to significantly dial up the disutility of dying or dial down the utility of killing psychopaths.

Of course, the Psychopath Button is a thought experiment, and all thought experiments involve abstraction. But we must remember that our intuitions are *our* intuitions and not those of some counterpart with vastly different psychologies to our own. So, insofar as many of us are non-expected utility maximisers or have psychologies vastly different to the agent in Egan's case, we ought to be cautious when insisting that Push is irrational for the agent Egan describes.

You might respond that we can set the credences in the Psychopath Button to $C(\text{Psychopath}) = \epsilon$, where ϵ is small enough such that given a reasonable utility (and perhaps risk-weighting) function EUT, REU, WLU, or whatever you like recommends pushing.

But how low would ϵ have to be, realistically? Very low, I suggest. Though we cannot say anything precise without filling in the details of the case, I would need to be *very* confident that I am not a Psychopath to push the button. What then could justify such a low credence? We might leverage the fact that very few of us are psychopaths—we infer from a population-level frequency (most people are not psychopaths) to a probability judgement about ourselves (I am likely not a psychopath). But statistical evidence is notoriously slippery. In order to apply statistical evidence to an individual case, we have to make background modelling assumptions: though $x\%$ of people are non-psychopaths, what about people in your reference class (philosophers, decision theorists, people contemplating mass murder)? Because of such complications, Steele (Manuscript) argues that the credences we are justified in forming on the basis of demographic statistics are typically more moderate than those statistics—if $x\%$ of people have trait y, your credence that some individual has trait y should be lower than $.x$. To get the ϵ-credence required for a realistic Psychopath Button, I claim we would need a lot more than a coarse-grained fact about overall frequencies of psychopaths.

Moreover, the mere fact that you are *considering* pushing might increase your confidence in Psychopathy. If only a Psychopath would push the button, what does it tell you that you are considering pushing? And if you feel slightly inclined to push (e.g., by learning that the causal expected utility of doing so is high), then this will increase your credence in your own Psychopathy even more. So, whatever evidence I have that I am not a psychopath must be robust enough to preserve my ϵ-credence when I learn that I really am considering pushing the Psychopath Button. This means that we would need something strong—psychological testing, a detailed knowledge of typical psychopathic traits, and so on, to get the verdict Egan objects to. But if I have that kind of evidence, how then can the mere fact that I push override that evidence and convince me that I am a psychopath? This puts us in an odd situation: for Egan's case to be realistic you need strong evidence that you are not a psychopath, but that evidence has to be fragile enough that your pushing is even stronger evidence that you are a psychopath. I struggle to think of what kind of evidence has this profile. Though not impossible, I suspect that in realistic cases (or at least the kinds of cases we can easily imagine when presented with Egan-cases) either (i) CDT will not recommend pushing (because you do not have strong evidence that you are not a psychopath), or (ii) pushing will not provide significant evidence that you are a psychopath (because you already had strong evidence that you are not a psychopath).

In sum, if Egan cases are possible, they are very strange. When I take seriously the credences and utilities involved in the Psychopath Button and begin to fill in the appropriate background conditions, I am not sure what my intuitions are. I am not even sure what such a case would look like. So, I am hesitant to place too much weight on my intuitions in Egan's case. Even if highly asymmetric cases of instability are possible, I doubt we can imagine them with the detail or consistency required to form a reliable intuition about them.

2.4.2.3 Reasoning About Correlations

Finally, we come to the elephant in the room: the non-causal correlations involved in Egan's case. The Psychopath Button asks us to imagine a strong statistical correlation without causation. To be sure, causation does not equal correlation (at least not by the two-boxers lights). But as is often pointed out, it does wink suggestively in its direction. Eriksson and Rabinowicz (2013) cite several studies which show that we conflate statistical with causal correlations. For example, Shafir and Tversky (1992) use the Newcomb Paradox to illustrate what they call 'quasi-magical' thinking: provided some outcome is unknown, people act as if their acts are causally efficacious, even if they know that there is no causal connection between act and outcome. Shafir and Tversky (p. 455) found the following pattern when presenting students with the Prisoner's Dilemma: students virtually unanimously defected when they were told that their opponent defected (choosing to cooperate in only 3% of games), students tended to defect when told that their opponent cooperated (choosing to cooperate in 16% of games), though a sizeable minority (37%) chose to cooperate when not told what their opponent had done. The fact that students cooperated more given conditions of uncertainty indicates that uncertainty is doing a lot of work. The most natural interpretation, suggested by Morris et al. (1998), is that when we do not know the outcome of a gamble, we treat it not only as unsettled but as under our control.[109]

This leads me to think that even those committed to a non-backtracking interpretation of counterfactuals may not be able to 'see' the Psychopath Button in non-backtracking terms. Even more than the standard Prisoner's Dilemma, Egan's case involves a strong correlation that is difficult to explain without causal dependence. The possibility of quasi-magical thinking should therefore make us hesitate about our intuitions in that case.

2.4.2.4 Dominance and Individuating Options

The causalist thinks that acts should be evaluated on whether they causally promote the good. From this perspective, there is a simple dominance argument for pushing the Psychopath Button. Consider:

> *The Two-Button Defence*: In front of you are two buttons. Each button kills all the psychopaths in the world, which you would like (though you value your own life more than killing all psychopaths). The first button, Button A, provides no evidence for psychopathy. The second button is just the Psychopath Button. The buttons differ in no other way that you care about (you are indifferent between survival and pushing A and survival and pushing the Psychopath Button; neither button causes you to become a psychopath). You must choose which button to press.[110]

We can model this with any values we like:

[109] Note that the evidentialist need not call this irrational. If students have the background belief that their acts are statistically correlated with their opponents', then EDT might say that quasi-magical thinking is rational. I am not sure whether people do operate with such a background belief when presented with the standard Prisoner's Dilemma (see Ahmed 2014b, Section 4.6 for helpful discussion).

[110] This kind of case is discussed by Ahmed (2012) who also comes to the conclusion that Egan-cases provide no reason to reject CDT (over and above the reasons mere instability already provides).

	Psychopath	Not Psychopath
Button A	x	y
Psychopath Button	x	y

Say that x, y are such that $\text{Button A} \succ \text{Status Quo}$. Now, this case does not directly say that you should Push over Refrain in Egan's original Psychopath Button. But it does show that not pushing the Psychopath Button while upholding Causal State-wise Dominance requires you to violate transitivity (see Ahmed 2012, Williamson 2021).[111]

But note that Briggs (2010b) provides an impossibility theorem showing that anyone who wishes to accept Egan's intuitions in the Psychopath Button must give up some core decision-theoretic principle—so perhaps transitivity might go in light of the Psychopath Button. (Moreover, as I will argue in Chapter 5, I am not convinced that Transitivity is even a desirable property, all else being equal.)

Here is what I now think the right way of thinking about the Two-Button Defence is.[112] We ought not individuate between acts that differ in merely superficial respects (for example, if the speed at which I push a button is irrelevant, then 'Pushing quickly' and 'Pushing slowly' count as the same act from the perspective of instrumental rationality). Though we might *say* that 'push quickly' and 'push slowly' are distinct acts in ordinary language, a theory of instrumental rationality ought to treat them as the one act. The formal objects that decision theory traffics in should not include extraneous information about properties that have no bearing on the efficacy of acts. For Savage, this is captured by the fact that each act is defined as a mapping from states to outcomes—two acts count as the same if they bring about the same outcome in each state of the world.

The causalist thinks that what matters is causally promoting the good, so call an act's *causal profile* a description of which outcome it brings about in each dependency hypothesis. Recall that each outcome describes everything of intrinsic concern to you, and each state is rich enough to specify appropriate act-outcome dependencies. So, from the causalist's perspective, a causal profile is a description of everything that matters when describing each act. Facts not included in the causal profile are therefore irrelevant.

[111] This argument is given in detail in Ahmed (2012): $\text{Button } A \sim \text{Psychopath Button}$ by Causal State-wise Dominance. Since $\text{Button } A \succ \text{Do Nothing}$ by definition, Transitivity requires $\text{Psychopath Button} \succ \text{Do Nothing}$.

[112] Thanks to Melissa Fusco for very helpful discussion here.

But if we buy that acts are individuated by causal profiles, then we are forced to say that Button A and the Psychopath Button are *the very same act*. So, any causalist who judges Push A to be better than the Status Quo, but does not say that the Psychopath Button is better than Refrain, does something worse than violate transitivity: they treat the same option as if it were two! Anyone attracted to causalism who takes Button A, but not the Psychopath Button, over the Status Quo is therefore guilty of a kind of incoherence that *nobody*, causalist or evidentialist, has defended: preferring an option to itself.

The Two Button Defence therefore raises the following challenge for the causalist: what account of options can you give that allows us to individuate between options with identical causal profiles? The causalist already thinks that causal efficacy is what matters and that outcomes capture everything you care about, so I do not see why they would treat acts that differ only with respect to evidential bearing differently. Any reason you have not to Push the Psychopath Button (by the causalist's lights) is a reason not to push Button A, and any reason you have to push the Button A (by the causalist's lights) is a reason to push the Psychopath Button. The only difference between the Psychopath Button and Button A is that one is stable and the other is not. So, the only reason to treat the Psychopath Button differently would be if we cared about the evidence an act provides—but that would be to pay attention to the backtracking counterfactual 'if I were to push, I would likely be a psychopath'. And the causalist has already reasonably rejected the relevance of such backtracking counterfactuals. So, the Two Button Defence highlights that there is *no difference* between the Psychopath Button and Button A; at least, there is no difference that the causalist has not already reasonably concluded is irrelevant. I do not think that a theory that deserves to be called 'causal' should individuate between options with identical causal profiles. This means that no theory that deserves to be called causal should treat 'Push the Psychopath Button' and 'Push Button A (which carries no news about your psychopathy)' as distinct acts.

2.4.2.5 Why Ain'cha Alive?

All well and good, but why ain'cha alive? What motivation is there to follow the causalist standard of rationality when the causalist ends up dead?

The causalist can simply give the same response as before: holding fixed facts outside your causal influence, the causalist does, in all probability, not end up dead. Recall the key premise in the WAR argument against Two-boxing:

1. The average return to One-boxing is higher than the average return to Two-boxing.

The causalist denies this premise because in the reference class of people with the same causal history as them, the average One-boxer does worse than the average two-boxer. Now, it is slightly more complicated to hold fixed facts causally independent of your decision in Egan's case since there are two ways the world could be, one leading to death, the other to survival. The reasonable thing is to consider the reference class of people who reflect your credences about yourself—most of them are non-psychopaths, and only a handful are psychopaths. Again, it is (likely) false that:

1*. The average return to pushing is lower than the average return to surviving.

That is, the causalist simply reasons, '*I* am unlikely to be a psychopath, so holding fixed facts outside my causal influence, it is very likely that *I* do better by pushing than refraining.' The average person in the reference class of people like you does better by pushing, so holding fixed facts outside your influence, you think that pushing foreseeably does better than refraining. Of course, *if* you decide to push then you end up in a new reference class in which people who push do worse; but that is just to say that your decision is unstable, and it does not change the fact that *now* pushing foreseeably does better than refraining. So, by the causalist's lights it is a reference class fallacy to assert that pushing does foreseeably worse than not pushing.

One last time: to consider the consequences of pushing by the lights of your views post-pushing is simply backtracking. And the causalist already rejects backtracking reasoning.

2.5 Conclusion: New Challenges, Old Responses

In this chapter, I have argued that the debate between causalist and evidentialist is at a stalemate. I have diagnosed this as arising from (i) the need to apply dominance reasoning only when states are independent of your act, and (ii) there being multiple legitimate standards of independence. I have shown how the initial disagreement about independence furnishes both causalist and evidentialist with the tools to address prominent arguments against their own view. This holds in the case of classic arguments (arguments from Full Information and Instability) as well as more sophisticated recent versions of those arguments (Briggs' Self-Sovereignty, Ahmed & Spencer's Actual Value, and Egan's Extreme Instability arguments).

Given the size of the debate between causalists and evidentialists, I have not chased every rabbit down every hole. But I hope the recipe for defusing debates is clear—pick an argument against causalism or evidentialism, look for where independence-talk is used (or appealed to implicitly), and check to see if the argument is question-begging. I predict that most arguments can be defused in this way. In the next chapter, I turn to an exception to this rule and consider a family of counterexamples that arguably undermines CDT even by its own lights.

Chapter 3

Determinism and Decision

3.0 Determinism and Doing Otherwise

The causalist asks what causally depends on their options. The evidentialist considers what the world would have been like for you to perform various options. I have argued that many standard arguments beg the question and so do not move us past the initial disagreement between causalist and evidentialist. I now turn to a family of challenges that does not fit this mould. Arif Ahmed (2013, 2014a, 2014b) has argued that *deterministic cases* provide us with reason to reject CDT; Ahmed's challenges do not lose their force if you already adopt the causalist framework, so they may break the stalemate.[113] I will focus on two cases, Betting on the Past and Betting on Laws and argue that the causalist can respond to both.

Before spelling out the challenges in detail, here is what I take to be the driving thought behind both cases. Because CDT reasons in terms of non-backtracking or causal counterfactuals, it asks what would happen were you to perform various acts, even if facts outside of your causal influence make it unlikely that you perform those acts. In this way, CDT severs the connection between your options and the past, the laws, and whatever else is outside of your causal influence. It treats you as unconstrained in the sense that it evaluates options that are unlikely (given the past, the laws, and so on) in just the same way as it treats options that are likely (given the past, the laws, and so on). This is what yields the two-boxing verdict in Newcomb's Paradox. But Ahmed aims to show that CDT treats you as unconstrained in an objectionable way—not only does CDT take seriously options that are unlikely given facts about the past, CDT takes seriously options that are *impossible* given some facts about the past. In particular, supposing you are determined not to do f, CDT asks what would happen if you were to do f and takes such facts seriously when working out what to do. This means that the causalist ends up acting as if they can break the laws of nature (or at least acting in a way that makes no sense if they cannot break the laws of nature). And this is absurd, even by the causalist's own lights.[114]

[113] Determinism itself is not a recent addition to the debate over Newcomb Paradox—determinism is discussed in Nozick (1969). What is recent is the way that determinism has been leveraged in several novel arguments by Arif Ahmed.

[114] Note Lewis' (1981b, p. 115) distinction between the plausible Weak Thesis, that 'I am able to do something such that, if I did it, a law would be broken', and the 'utterly incredible' Strong Thesis, that 'I am able to break a law'. One way of framing Ahmed's challenge is that if we commit to the Weak Thesis and CDT, we thereby accept (something like) the Strong Thesis.

My goal in this chapter is to outline a view that lets us reason causally without treating ourselves as free in this objectionable way. The view I end up defending will not be CDT but a close variant, *Selective* Causal Decision Theory (SDT), which I think captures enough of the motivation behind CDT to be called causal. The motivating thought behind SDT is that you should causally promote the good, which means ignoring statistical correlations between acts and outcomes, while taking seriously the constraints that laws of nature place on you.

Here is the plan. I outline the first deterministic challenge, Betting on Laws, and argue that it is not a problem for CDT. Betting on Laws only appears to be a problem because our ordinary semantics for counterfactuals is not well-suited to cases involving free-will and determinism. I argue that an impossible worlds semantics for counterfactuals, building on Nolan (1997, 2017), puts CDT on a sure footing.[115] I then outline the second deterministic challenge, Betting on the Past, and argue that this case is indeed a problem for CDT. I propose a distinction between those counterfactuals that are worth taking seriously and those that are not worth taking seriously for the purposes of practical deliberation. This distinction allows us to treat Betting on the Past differently from familiar Newcomb Problems. SDT agrees with CDT in most cases but delivers the correct verdict in Betting on the Past.[116] I conclude that a broadly causalist position is coherent despite *prima facie* counterexamples involving determinism.

3.0.1 Counterfactuals, Causation, and Decision

Recall that CDT is primarily a thesis about how to define the states in a decision situation. By defining states as causal dependency hypotheses, we have followed Lewis in embedding causal concepts into decision theory to get the two-boxing verdict in Newcomb's Problem. The first challenge I discuss is clearest, however, if we adopt an alternate formulation of CDT in terms of counterfactuals. So, rather than working with:

$$U_C(a) = \sum_k C(k) \cdot u(o_{a,k})$$

I work initially with:

[115] This section draws on co-authored work with Alexander Sandgren, 'Law-Abiding Causal Decision Theory' (Forthcoming).

[116] This section draws on co-authored work with Alexander Sandgren from our 'Determinism, Counterfactuals, and Decision' (2021).

$$U_{CF}(a) = \sum_k C(a \to o_{a,k}) \cdot u(o_{a,k})$$

It is important to note that this explicitly counterfactual formulation of CDT is supposed to be a repackaging of our original Lewisian formulation (modulo concerns about probabilities of conditionals, conditional probabilities, and how this all relates to causation). Some prefer a statement of CDT in terms of credences in counterfactuals, rather than dependency hypotheses (e.g., Gibbard & Harper 1978, Hedden Manuscript). But here I take the causalist standard of rationality to be U_C-maximisation, with U_{CF} being a repackaging of U_C. (For the purposes of this chapter, every reference to → is a causal, non-backtracking counterfactual.)

3.1 Lawless Decision Theory

Ahmed (2013; cf. 2014b) argues that CDT goes wrong in the following case:

> *Betting on Laws*: You are a scientist about to publish a ground-breaking paper on whether the universe is governed by some system of deterministic laws, L. The paper will contain several strong theoretical arguments in favour of L as well as the results of numerous experiments confirming L. The only job left for you is to write the conclusion of the paper, in which you have two options: *endorse* L (call this o_1) or *deny* L (call this o_2). You only care about truth, such that the outcomes are:

	L	$\neg L$
o_1	Win	Lose
o_2	Lose	Win

And let us stipulate that $u(\text{Win}) = 1, u(\text{Lose}) = 0$.

Given this setup, you clearly ought to endorse L. The payoffs are symmetric, so you should simply endorse whichever proposition you are more confident in, and you are *overwhelmingly* more confident in L than its negation. From here, I assume that a 'Win' holds at a world if and only if that world is an $(o_1\&L)$- or $(o_2\&\neg L)$-world; similarly, I assume that 'Lose' holds at a world if and only if that world is an $(o_1\&\neg L)$- or $(o_2\&L)$-world.

The problem is that CDT seemingly permits you to deny L. I reconstruct Ahmed's argument (2013, pp. 294-5) as follows:

Premise 1: $o_1 \rightarrow \text{Win}$ entails $o_2 \rightarrow \text{Win}$.

Premise 2: If $o_1 \rightarrow \text{Win}$ entails $o_2 \rightarrow \text{Win}$, then $C(o_2 \rightarrow \text{Win}) \geq C(o_1 \rightarrow \text{Win})$.

Premise 3: If $C(o_2 \rightarrow \text{Win}) \geq C(o_1 \rightarrow \text{Win})$, then CDT permits o_2.

Conclusion: CDT permits o_2.

Premise 1 follows from some basic claims about the semantics of counterfactuals and the nature of determinism. Firstly, Ahmed (2013, p. 290) adopts Lewis' (1979, p. 460) characterisation of determinism: if two possible worlds are governed by the same deterministic laws, then if those worlds agree with respect to matters of particular fact at a time, they agree with respect to matters of particular fact at all times. Determinism guarantees that we get a complete tape of the world simply from pressing 'fast forward' or 'rewind' at a single point in time—the laws fully constrain everything that happens from that point in time.[117] We can now derive **Premise 1** from two sub-premises about counterfactuals:

Premise 1a: If you do act a and L is true, then $\neg a \rightarrow \neg L$ is true.

Premise 1b: If you do not do act a and $a \rightarrow L$ is true, then $\neg L$ is true.

Both of these sub-premises follow if we accept the standard Lewisian story, on which $X \rightarrow Y$ is true if the closest X-worlds are Y-worlds, where the closest X-worlds are those that diverge from our own due to a small pre-X miracle.[118] To see that **Premise 1a** holds, say that $a\&L$ is true. Since L is deterministic, then L along with facts about the past determines that you do a. The nearest $\neg a$-worlds are therefore worlds that match our own until a small miracle just before a would have occurred. Since such worlds match the actual world at pre-miracle times, they cannot be L-worlds by our characterisation of determinism. So, the closest $\neg a$-worlds are $\neg L$-worlds, giving $\neg a \rightarrow \neg L$. **Premise 1b** holds for similar reasons. If you do not do a and the nearest a-worlds are L-worlds, then those nearest $a\&L$ worlds match our own at pre-miracle times,

[117] Strictly speaking this is a two-way or time-reversible characterisation of determinism. I set aside the possibility that we might be able to fast forward the tape of the world but not re-wind (say, if we can fast forward to the heat death of the universe but, because multiple worlds end in the same heat death, cannot re-wind the tape from that point).

[118] This is certainly not the only possible account of counterfactuals—I will discuss others in Sections 3.1.3 and 3.1.4.

meaning that if L is true at our world you also do a in our world. But you do not do a in our world, so the actual world is a $\neg L$ world.

Having established **1a** and **1b**, note that $o_1 \rightarrow \text{Win}$ can be true in one of two ways:

(1) o_1 & $(o_1 \rightarrow \text{Win})$
(2) $\neg o_1$ & $(o_1 \rightarrow \text{Win})$

First take (1). Since o_1 holds, the actual world is o_1. Therefore, because $o_1 \rightarrow \text{Win}$ holds the actual world must be an L-world (simply because the closest o_1 world is the actual world, hence a Win-world since in it you endorse L and L is true). So we have an instance of the antecedent of **1a**, namely that you do o_1 and L is true. So, we can conclude that $o_2 \rightarrow \neg L$, meaning $o_2 \rightarrow \text{Win}$ (the closest worlds in which you deny L are worlds in which L is false, hence worlds in which you win your bet). Next take (2). This is an instance of the antecedent of **1b**—you do not do o_1 and $o_1 \rightarrow L$ is true. So the actual world is a $\neg L$-world. So $o_2 \rightarrow \neg L$ is trivially true, hence $o_2 \rightarrow \text{Win}$. So in both cases (1) and (2), $o_1 \rightarrow \text{Win}$ entails $o_2 \rightarrow \text{Win}$, this concludes the argument for **Premise 1**.

Premise 2 follows straight from the probability calculus. **Premise 3** follows by noting that both:

$$U_{CF}(o_1) = C(o_1 \rightarrow \text{Win}) \cdot 1 + C(o_1 \rightarrow \text{Lose}) = C(o_1 \rightarrow \text{Win})$$

$$U_{CF}(o_2) = C(o_2 \rightarrow \text{Win}) \cdot 1 + C(o_2 \rightarrow \text{Lose}) = C(o_2 \rightarrow \text{Win})$$

So, CDT permits you to deny L. And note that this argument makes no reference to your credence in L. You can be as confident as you like in L and CDT will still permit you to deny it. But this is wrong—the scientists who is overwhelmingly confident in some deterministic proposition (e.g., that Newtonian physics is true) ought to endorse that thesis.

The problem is that on the Lewisian account of counterfactuals, deterministic propositions are 'modally fragile'. The standard Lewisian view licenses the following kind of reasoning: 'Say that L is true and I am determined to endorse L—I would do no worse by denying L (since L would have to be false for me to deny it)! Likewise, say that I am determined to deny L—I would do no better by endorsing L (since L would have to be false for me to endorse it)!'. If some deterministic proposition is true, then it would be false were you to do otherwise. CDT takes this modal fragility as a reason not to endorse deterministic theses and so delivers the wrong verdict in Betting on Laws.

3.1.1 A Lewisian Worry

To formulate a response to Betting on Laws, note that CDT goes wrong based on a U_{CF} calculation. But, as emphasised above, this is a particular interpretation of CDT. So, we ought to ask whether we can get the same verdict simply by calculating U_C. To do so requires us to specify the dependency hypotheses in Betting on Laws.

What then are the dependency hypotheses in Betting on Laws? Forget for a moment that you have ever thought carefully about the semantics for counterfactuals and the modal fragility of deterministic propositions. If you simply read the case, you would likely think the obvious candidate for a partition is $K = \{L, \neg L\}$. Call this *the L-partition*. Note how natural it was to follow Ahmed (2013, p. 291) in using the L-partition to specify the columns of our decision table above. And if we calculate U_C in line with this partition, we get:

$$U(o_1) = C(L) \cdot 1 + C(\neg L) \cdot 0 = C(L)$$

$$U(o_2) = C(L) \cdot 0 + C(\neg L) \cdot 1 = C(\neg L)$$

And this says you should endorse L whenever you are more confident in L than its negation—the correct verdict! This is puzzling. We get two different verdicts when we use U_C and U_{CF} respectively, though they are supposed to be statements of the same theory. What is going on?

The solution is that the U_{CF} calculation implicitly requires us to reject the L-partition.[119] This raises some important questions: which is the correct partition to use, the L-partition or the partition implicit in the U_{CF} calculation? And is there a way of spelling out the contents of the L-partition such that its members qualify as dependency hypotheses? I think the causalist can both motivate the L-partition and do so in a way that does not require us to reject a counterfactual analysis of causal dependence.

3.1.2 Motivating the L-Partition

[119] Instead, it assumes the dependency hypotheses are $C_1 = \{o_1 \rightarrow \text{Win}, o_2 \rightarrow \text{Lose}\}, C_2 = \{o_1 \rightarrow \text{Lose}, o_2 \rightarrow \text{Win}\}$, with these counterfactuals interpreted in line with the standard Lewisian semantics.

If the members of the L-partition are appropriate candidates for dependency hypotheses, then CDT gets things right. So, the question becomes whether L and $\neg L$ are appropriate candidates for dependency hypotheses. I think that they are. Firstly, note that everyone thinks that the laws are causally independent of your choice and outside of your causal influence.[120] Moreover, it is the truth (or falsity) of L that determines the outcome of your choice. If L is true, then it is in virtue of this fact that you win by endorsing L. Similarly if L is false, then it is in virtue of this fact that you lose by rejecting L. *These* are the commonsense judgements underwriting our intuitions in Betting on Laws, and it is these judgements that our decision theory should respect. The truth of L is causally independent of your choice and tells you how things you care about do and do not depend on your choice in an intuitive and plausible way. And dependency hypotheses are just that—propositions that are causally independent of your choice and that tell you how things you care about do and do not depend on your choice. So, the members of the L-partition are good candidates for dependency hypotheses.

But simply insisting on the L-partition is not by itself a complete response on behalf of the causalist. It leaves us without a general recipe for formulating dependency hypotheses, and a significant advantage of the counterfactual approach is that it provides us with such a recipe. Moreover, it leaves us without any precise way of saying what the relationship between act and outcome is and the sense in which your winning (losing) depends on your endorsing (denying) L. We can say more.

3.1.3 Doing the Impossible

CDT goes wrong when coupled with Lewis' theory of counterfactuals because L is modally fragile: if L determines that you do a, then L is false at nearby possible worlds in which you do not do a. But you might just think that this is a problem with Lewis' 'small miracle' account of counterfactuals. You cannot causally influence the laws and yet the standard semantics says that your doing otherwise would involve the laws being different. Strange indeed. Moreover, we have to block the inference from 'if I were to do otherwise the laws would be different' to 'the laws

[120] Well, perhaps not *everyone*. On a Best Systems Analysis of laws, you might be a part of making some laws hold—say, by ensuring that a regularity holds. If this does entail that you should bet against the best confirmed theories in your physics textbook, then this is a reason to reject the Best Systems Analysis.

depend on my acts'. But if we adopt a counterfactual analysis of causal dependence, it seems natural to analyse the latter in terms of the former.

One route would be to sever the connection between counterfactuals and causation. But there is another. While your doing otherwise might involve the laws being different at nearby possible worlds, some are already convinced that we need *impossible* worlds to make sense of the relationship between action and the laws of nature. In particular Nolan (2017) claims that if we take the possibility of determinism seriously, we need impossible worlds to make sense of certain plausible counterfactual judgements. He argues:

> When considering a counterfactual situation, we initially assume the same fundamental principles of nature are at work … for most of these antecedents [i.e. antecedents of counterfactuals involving things going differently from the way they actually do], the relevantly similar worlds where they are true are not ones where the laws of nature differ from the actual laws. (Nolan 2017, p. 19)

I agree. While counterfactuals involve things going differently, those differences do not generally go all the way down to the laws of nature. My dropping a cup might make a difference to whether the cup breaks, but not to whether Newtonian physics is true. Of course, we could keep the laws the same if we were to tweak the distant past (see Dorr 2016). But it also sounds wrong to say, 'Had I dropped the cup, the initial stages of the Big Bang would have been different' (more on this and Dorr's solution soon). But then we have a puzzle: if the laws along with the past determine that I do not drop the cup, then how could I drop the cup without different laws or a different past? This is Nolan's trilemma (2017, Section 3): tweak the laws, tweak the past, or give up on the possibility of doing otherwise.

Impossible worlds resolve the trilemma.[121] We say that even if some set of laws and past facts determines that you do a, there are worlds with those laws and past facts in which you do not do a.[122] Instead of making the laws or past dependent on your choice, we enrich our conceptual toolbox with impossible worlds.

[121] Let me forestall one objection—impossible worlds, *really* (perhaps followed by an incredulous stare)? Since this is not the place to weigh the pro's and con's of impossible worlds, I can only say that I find them plausible enough. For the purposes of this chapter, we can think of worlds (possible or impossible) as modelling tools—objects that we introduce to capture ways things might go, so that we can then do useful things like assess the truth of counterfactuals. You might like to think of worlds as formal objects, sets of sentences, sets of propositions, stories, and so on. An implicit goal of this chapter is to show that whatever worlds are, it is fruitful to be able to represent the impossible.

[122] Introducing impossible worlds may not *resolve* the dilemma so much as sweeten the third horn. Even if it is in some sense impossible for agents to do otherwise, impossible worlds still allow us to make sense of what would happen if they did otherwise.

If we want ordinary claims like 'If I were to drop the cup, the past and the laws of physics would be the same' to come out true, we have to say something about an impossible worlds semantics for counterfactuals. In particular, why should the closest worlds, whether possible or impossible, be ones with the same laws and past? One thing we must deny, following Nolan (1997, p. 550) is:

> **Strangeness of Impossibility Condition:** Every possible world is closer to every other possible world than any impossible world is to any possible world.

This condition would prioritise possible worlds in our similarity ranking, meaning that impossible worlds are too distant to affect the truth of ordinary counterfactual claims. But Nolan's trilemma shows that impossible worlds may be necessary to get ordinary counterfactuals to come out as true, so we should reject the Strangeness of Impossibility Condition (cf. Nolan 2017, p. 29).

We can now retain the core of Lewis' semantics: $X \rightarrow Y$ is true if and only if the closest X-worlds are Y-worlds (where the closest X-worlds might be possible or impossible). The crucial thing then is that we prioritise match with respect to the laws and past matters of fact—your doing otherwise would not involve different laws or a different past. Now, settling on the precise details of a similarity ranking is a vexed issue, and it would go well beyond the scope of this thesis to defend such a similarity ranking in full. What matters is that the correct similarity ranking ensures that the laws and the past are modally robust, at least under your doing otherwise. This suggests one further condition:

> **Localised Impossibility:** If the actual world is possible, then pro tanto, the more law-violations there are at world w, the less close w is to the actual world. ('Pro tanto' here means at least that if a counterfactual's antecedent occurs at time *t*, then minimising law-violations trumps securing match with respect to particular matters of fact from *t* onward.)[123]

Nolan does not defend such a condition explicitly, but he does articulate a good reason to accept it:

> The 'explosion' world—the impossible world where every proposition is true—is very dissimilar from our own … On the other hand, the world which is otherwise exactly like

[123] This raises an interesting question that I cannot fully address here: how should we individuate law-violations? Here one way of doing so: a law-violation occurs in w at time t if (i) w is an L-world, (ii) d is a complete description of the matters of particular fact at t in w, but (iii) w does not evolve from t in accordance with d and L. So, one implication of Localised Impossibility is that a world is closer if there are fewer times at which the world does not evolve according to the laws.

> ours, except that Hobbes succeeded in his ambition of squaring the circle (but kept it a secret), is far less dissimilar. (Nolan 1997, p. 544)

Localised Impossibility guarantees that the explosion world is indeed more distant from our own—it has far more impossible facts and events (the impossibilities are extremely non-local). The world in which Hobbes squares the circle, however, contains a localised impossibility and so is comparatively close to our own. Similarly, a world in which a neuron misfires once (violating the laws) is more similar to ours than a world in which people regularly fly faster than the speed of light. Moreover, if we prioritise match with respect to particular matters of fact, worlds at which things evolve lawfully until t and then violate the laws in some small region of space are closer to ours than a world in which the laws are violated in large regions of space. So, a world that involves (i) an unlawful neuron firing at t but at t things otherwise being as they are in the actual world, is closer to our own than (ii) a world in which large, wide-spread regions of space do not evolve lawfully.

Generally (see Nolan 1997, p. 544) we want worlds where I do otherwise to diverge from our world. For example, given that I just dropped a glass, we want 'If had not dropped the glass, it would (very likely) not have shattered' to come out true, even though the glass does shatter in the actual world. Without Localised Impossibility, such ordinary counterfactuals might not come out as true. For any counterfactual involving things going differently, we could simply add some future law-violations and return things to the way they are in our world—'If I were to drop the glass, there would be another law-violation and it would not shatter'. So, just as Lewis wants nearby possible worlds to have *similar* laws to ours (no miraculous non-shattering glasses in nearby miracle-worlds), we want counterfactual worlds to minimise law-violations (no unnecessary law-violations).

Where does this leave us? The closest worlds in which you do otherwise (i) have the same laws and past as our own, (ii) diverge at the moment of your doing otherwise, which might involve a small, localised impossibility (e.g., a neuron misfiring), and (iii) evolve lawfully from that time on. A heuristic is to think of the relevant impossible worlds as two possible worlds 'glued together' by a local impossibility. Your doing otherwise would involve things going precisely as they did given the actual laws and past, your acting in a different way at *t* (even though your doing so is impossible given the actual laws), and the universe subsequently evolving in accordance with the actual laws and matters of particular fact at *t*. The localised impossibility here plays the role of Lewis' small miracle, though the laws remain the same.

This impossible worlds framework puts us in a position to solve Betting on Laws without abandoning a counterfactual formulation of CDT. In the impossible worlds framework, $o_1 \& L$ no longer entails $o_2 \rightarrow \mathsf{Win}$. If L is true, then the closest worlds in which you endorse (deny) L are worlds in which L is true and so worlds in which you win (lose) your bet.[124] That is:

$$o_1 \rightarrow \mathsf{Win} \text{ if and only if } L$$

$$o_2 \rightarrow \mathsf{Win} \text{ if and only if } \neg L$$

So we get the L-partition: $L = \{o_1 \rightarrow \mathsf{Win}, o_2 \rightarrow \mathsf{Lose}\}, \neg L = \{o_1 \rightarrow \mathsf{Lose}, o_2 \rightarrow \mathsf{Win}\}$, where these counterfactuals are evaluated in line with an appropriate impossible worlds semantics.

Note that the impossible worlds semantics defuses an argument from Ahmed that any theory deserving to be called *causal* must permit you to deny L. Ahmed (2013, p. 301) introduces the following plausible principle:

> **Weak Causal Dominance:** For options a and b, suppose that any a-world w_a is at least as good (for you) as any b-world w_b that matches w_a over all matters of particular fact that are causally independent of your choice between them. Then it is rational for you to realize a when b is the only alternative.

This is plausible—if a-worlds are always better than b-worlds when those worlds agree with respect to all facts that hold independently of what you do, then a is at least as good as b. If the causalist did not endorse *this* thought, then I agree that it is unclear in what respect they would be a causalist.

Ahmed shows (2013, p. 301) that o_2 weakly causally dominates o_1 given that all worlds are possible worlds. But this claim does not hold in the more general impossible worlds framework. There are o_1-worlds that agree with o_2-worlds with respect to matters of particular fact outside your causal influence *and* that are better for you than those o_2 worlds—witness the impossible world in which L determines that you do o_2 but you do o_1. So, adding impossible worlds to the framework lets us respect Weak Causal Dominance in Ahmed's case. Insofar as Weak Causal

[124] Moreover, the closest worlds in which you win your bet are *only* worlds in which you win your bet, and the closest worlds in which you lose your bet are *only* worlds in which you lose your bet. You might worry that the closest worlds in which you endorse L may be worlds in which you (impossibly) deny L and so worlds in which you both win *and* lose (indeed, these worlds may increase match with respect to matters of particular fact from the time of your bet onward). To block this worry, we may stipulate that the closest X-worlds are not also $\neg X$-worlds. Moreover, Localised Impossibility is intended to rule such situations out—such a world multiplies impossibilities (the laws are violated *and* you perform contradictory acts) in order to secure match with respect to particular matters of fact. Thanks to an anonymous referee for discussion on this point.

Dominance is an intuitive constraint on the relationship between causation and choice, impossible worlds let us uphold an intuitive connection between causation and choice.

None of this should be surprising. There is a striking resemblance between Ahmed's case and Nolan's initial motivation for analysing causal counterfactuals with impossible worlds. Nolan is concerned with the strangeness of judgements like 'had I acted differently, the laws would have been different'—the modal fragility of the laws at the linguistic level. Betting on Laws is problematic because of precisely this fragility. We might even say that Betting on Laws is the decision-theoretic analogue of the linguistic problem that Lewisian semantics generates for counterfactuals. Just as it is foolish to assert that the laws would have been different had you acted differently, it is foolish to act differently so that the laws would have been different. Adding impossible worlds saves us from both kinds of folly.[125]

3.1.4 Why Go Impossible?

The key to solving Betting on Laws is to secure the modal robustness of the laws. But you might worry that impossible worlds are a heavy-handed way of achieving this. Or you might have independent worries about the coherence of an impossible worlds framework. Perhaps then we should look for a less radical solution to Betting on Laws?

I am happy if there is a less radical solution. The impossible worlds framework is intended to highlight precisely where Ahmed's challenge comes from (modal fragility) and provide a plausible solution. There may be other solutions that, depending on your background commitments, are more appealing. Here I sketch two possible routes, a no-miracles account of counterfactuals from Dorr[126] and a rigid designator strategy. I conclude that Dorr's solution fails, while the rigid designator strategy incurs costs that may be payable.

Firstly, Dorr (2016) defends a view on which 'had I done differently, the laws would be the same' is true.[127] Dorr avoids impossible worlds by allowing the past to vary at worlds in which

[125] This thesis is not the place to weigh the balance between impossible worlds' virtues and vices. Interested readers can consult Nolan (1997) for background. Bernstein (2016) discusses another application of impossible worlds to causal counterfactuals, those involving omissions. Note that I have been following Nolan in thinking that if in some world the laws determine a and a does not hold, then that world is impossible. Some might prefer a view on which laws can admit violations—if so then you might prefer the term 'miraculous possible world' for what I have been calling an impossible world.

[126] Thanks to an anonymous referee for pushing the relevance of Dorr's approach.

[127] The differences between Dorr and other same-past/no-miracles accounts of counterfactuals, such as Goodman (2015), need not concern us here.

you do otherwise. Since such a semantics could underwrite the L-partition, is this not a simpler solution to the impossible worlds strategy (see Dorr 2016, pp. 274-6 for a discussion of Betting on Laws)?

We might reply, as Nolan does (2017, Section 3.2) that it is odd to say 'Had I dropped the cup, the distant past of the universe would have been different', perhaps just as odd as it is to say 'Had I dropped the cup, Newtonian physics would be false'. Both claims place something that is outside of your causal influence under your counterfactual influence.

Dorr responds (p. 252) to this kind of worry by noting that the past differences required on his account involve highly specific facts (for example, the precise arrangement of atoms), and it is not absurd to say that *those* facts would have been different. After all, people spend entire careers trying to pin down the laws of physics—we hope that they can succeed! Not so with the precise arrangement of atoms yesterday. So, Dorr claims, the robustness of the past is less important than the robustness of the laws.

We might dissent from Dorr's linguistic judgements—the distant past is a long way away, and I bear no ill will towards the Laplacian scientist seeking to fully describe the earliest stages of the universe. And even if the robustness of the laws is *more* important than the robustness of the past, we might also judge the robustness of the past to be important. To say that the arrangement of particles shortly after the big bang would be different had I done differently is a cost, even if it is not as great a cost as letting the laws depend on my decisions. All else being equal, I prefer an account that secures the robustness of the past and laws to one that only secures the robustness of one.[128]

There is, however, a deeper worry with Dorr's approach: you are still forced to bet against your credences. Consider the following:

> *Betting on History*: You are about to publish a ground-breaking paper on whether the universe was ever in some particular state (call the proposition that the universe was at some point in this state 'H'). The paper contains several strong arguments in favour of H and the results of numerous experiments confirming H. Your own credence in H is very high (only just short of 1). The only job left is to write the conclusion of the paper, in

[128] Dorr (Section 3) argues that not all else is equal and that there are costs to accepting something like an impossible worlds account. I accept the various defences of impossible worlds offered in the literature. If Dorr is right, then I actually think Nolan's trilemma gives us a good reason to reject Causal Decision Theory: there is no sensible way of securing the modal robustness of all that matters, and so we should reject a theory that employs causal counterfactuals.

which you must either *endorse* H (q_1) or *deny* H (q_2). You are also convinced that determinism is true.

Betting on History is structurally similar to Betting on Laws, and the same issues arise. You are overwhelmingly confident in H and so ought to endorse it.

Dorr's approach, however, permits you to deny H. To see this, simply re-run the argument from Betting on Laws. If $C(q_2 \rightarrow \text{Win}) \geq C(q_1 \rightarrow \text{Win})$, then CDT permits you to deny H. And we get this inequality because $q_1 \rightarrow \text{Win}$ entails $q_2 \rightarrow \text{Win}$. Simply note that the analogues of **1a** and **1b** when we replace L by H for parallel reasons, given that determinism is true. So, CDT coupled with a no-miracles view tells you to deny some proposition regardless of your credence in it—your behaviour is insensitive to your credences. And no decision theory should tell you to bet against your credences. Even if Dorr's semantics captures plausible judgements at the linguistic level, it cannot play the decision-theoretic role that an adequate semantics should play, at least by the causalist's lights.

Might someone reply that Betting on History is less worrying than Betting on Laws because it is more far-fetched? I do not think so. CDT recommends that you do the irrational, and the correct decision theory never recommends the irrational. Moreover, causalists are already motivated by the desire to have a decision theory that handles far-fetched cases (Newcomb's Paradox is pretty far-fetched to begin with). And the case is not necessarily that far-fetched: we need not be able to describe H in detail to bet on it (we could just pick out H by saying 'the state that well-confirmed theory T says the universe must have been in at time *t*', or the like). There is nothing incoherent about Betting on History, so we ought to be able to say something sensible about it. The impossible worlds proposal makes the laws *and* past modally robust, so gives the correct verdict where Dorr-style approaches do not.

A second approach would be to introduce rigid designators into the description of cases involving modally fragile propositions. The modal fragility problem shows that we care about how the laws are *in our world*, not merely nearby worlds. So perhaps we should interpret Betting on Laws as a bet on the proposition that the actual world, rigidly designated, is an L-world. Let $L_@$ denote the proposition that @ (the actual world) is an L-world. What if we simply interpret Betting on Laws as a bet on $L_@$?

If we do so, then note that if $L_@$ is true, it is true at nearby possible words that are not L-worlds. Even though L is false *there*, those worlds agree that @ is an L-world. So, if $L_@$ is true, then you would win your bet on $L_@$ were you to bet on it, regardless of what L determines. Therefore, on any plausible semantics for counterfactuals:

$$o_1 \rightarrow \text{Win if and only if } L_@$$

$$o_2 \rightarrow \text{Win if and only if } \neg L_@$$

So $K = \{L_@, \neg L_@\}$ is an appropriate set of dependency hypotheses.

Should we simply endorse a rigid designator solution to Betting on Laws and sidestep impossible worlds? There are at least four reasons why you might reject this approach:

1) Rigid designators are odd, and they interact with counterfactuals in an odd way. Say that $L_@$ is true. Consider your counterpart who is in a $\neg L$-world and who denies $L_@$. On the rigid designator account, we have to say that they *lose* their bet on the proposition that 'the actual world is $\neg L$'. But the world they pick out as actual is a $\neg L$ world! What could you say to your counterpart: 'Sorry—you bet on the actual world being $\neg L$ and the world you pick out as actual is $\neg L$—too bad you didn't pick out the right world as actual'? Surely they would respond: 'I said the actual world is $\neg L$, and I'm right, so pay up!'? This is an odd situation for your counterpart.[129] Perhaps it is an oddness we can live with. So long as *we* can pick out @ with the use of an indexical (*this* world), and so long as our decision theory gives the right verdicts when we do so, then we might simply ignore your counterpart's quizzical look. If we think of possible worlds as tools to help us get the right results in decision theory, then we need not take your counterparts quizzical stare as an objection, provided our theory gets the right results. Of course, philosophers of language have a range of theories about rigid designation and a range of views about indexicals. If you are suspicious of the work this strategy puts them to, then this will push you towards the impossible worlds solution.
2) We need non-trivial credences in propositions like $L_@$. But note that @ rigidly designates a world, then $L_@$ seems to be true at *all* worlds (you might think that $L_@$ is *too* modally robust). And standard probability theory says that if a proposition is true (false) at all worlds, then it gets probability 1 (0). So, we need a probability theory that can assign

[129] Thanks to Daniel Nolan and an anonymous referee for raising this point.

non-trivial probabilities to necessary propositions.[130] If you are suspicious of this project, then that is a reason to employ an impossible worlds solution.[131]

3) Even if Betting on Laws is best interpreted as a bet on $L_{@}$, why can we not introduce a case and stipulate that it is a bet on L not rigidly designated (call this BETTING ON LAWS*)?[132] We might be able to say that a bet on L, on a possible worlds semantics for counterfactuals, is not really a legitimate decision. After all, L is not counterfactually independent of your decision, so the partition $K^* = \{L, \neg L\}$ is not act-state independent.[133] But if you think that we ought to be able to bet on the laws whether rigidly designated or not, then this might be a reason to prefer the impossible worlds solution.
4) We can rigidly designate anything we like. What about a bet on the proposition 'I do not actually raise my hand', where I get $1 if I do not actually raise my hand and $10 if I actually do raise my hand? Complication: to accept this bet, I must raise my hand. In nearby possible worlds in which I raise my hand, it might be true that I *actually* do not raise my hands. Perhaps we can deal with complications like this (see Williamson & Sandgren Section 6). Or perhaps you would rather not live with them, and this might be a reason to prefer the impossible worlds solution.

In sum, we may be able to motivate a possible worlds solution along with an adequate theory of rigid designation, a hyperintensional probability theory, and various caveats. I am not sure if the resulting theory will be any more parsimonious than the impossible worlds approach. What matters here is that we *can* solve Betting on Laws by guaranteeing the modal robustness of the laws. The problem therefore is not with CDT, but with the standard framework for thinking about causation and counterfactuals.

3.2 Betting on the Impossible

[130] Another approach would be to have @ pick out the actual world at a range of nearby worlds but not all worlds—something like a 'robust designator'. I am unsure how viable a robust designator approach would be.

[131] Note that on the impossible worlds solution, we need not assign positive probability to any impossible world. We simply set the credence of $X \rightarrow Y$ to be the probability of N, where N is the set of possible worlds at which $X \rightarrow Y$ is true. Some of the possible worlds we assign positive probability to will be located *nearby* impossible worlds, but we need not assign positive probability to those impossible worlds directly. Thanks to Alan Hájek for discussion here.

[132] Thanks to Boris Kment for discussion on this point.

[133] The details of this are spelled out in Williamson & Sandgren Section 5.2.1.

So far, I have argued that CDT is compatible with your betting on deterministic theses. But in defending CDT on this front, some may worry that I have taken CDT out of the frying pan and put it into the fire. CDT seems to take impossibilities *too* seriously, so some have argued.

To see the tension between CDT and determinism, consider the following:

> *Simone*: Simone is convinced that the universe follows deterministic laws. She is asked if she would like to bet on the truth of some proposition H about the past, to which she assigns credence $.99$. If H is true and she bets, then Simone wins a car; if H is false and she bets, then Simone loses $\$10$. Being a genius, Simone knows that H together with the laws of nature determines that she will not bet. What should Simone do? Clearly, Simone should not take the bet. After all, she knows that winning the bet would involve a law-violation, and Simone understands enough metaphysics to know that she is not free to break the laws.

This sketch of Simone's situation needs filling out in several ways. But here I want to highlight that Simone will go badly wrong by employing a particular kind of counterfactual reasoning. Suppose she reasons: If H is true, then I would win a car if I bet. If H is false, then I would lose $\$10$ if I bet. I am very confident that H is true and that I cannot affect H, so the small probability of losing a measly $\$10$ is offset by the high probability of gaining a car, were I to bet. Therefore, I should bet.

This is bad reasoning since Simone can only lose $\$10$ by betting. So, the above (non-backtracking) counterfactual reasoning is not a good guide to choice. But counterfactual reasoning lies at the heart of causalism, as I have spelled it out. CDT tells you to think about the various things you could do and asks what *would* happen were you to do them; this screens off information about the predictor's accuracy in Newcomb's Paradox, but in Simone's case it also seems to screen off relevant information about the past. Reason counterfactually and ignore act-state correlations—so the causalist says. But if you reason counterfactually and ignore deterministic connections, you doom yourself to failure. This tension will create serious problems for CDT.

My core claim is this: Simone's case does not show that counterfactual reasoning is irrational. Rather, Simone's case shows that there are particular *kinds* of irrational counterfactual reasoning. Causalists take Newcomb's Paradox to show that non-backtracking counterfactuals are important: you should evaluate acts based on causal influence, not their evidential bearing. Mere correlations between choices and the past do not prevent you from thinking of yourself as free and reasoning about what would happen if you did otherwise, holding fixed the past. We need a principled reason for Simone to avoid the counterfactual reasoning that led her to bet on H while allowing her to reason counterfactually in Newcomb's Paradox.

Here is such a principled reason. Simone is very confident that the following counterfactual is true: 'if I were to bet, I would win a car'. But that counterfactual is *irrelevant* to what Simone should do. Why? Because she knows that if H is true, then the laws determine that she does not bet. So, worlds in which she wins her bet must have different laws to ours. It is therefore futile for Simone to bet on H in the hope of bringing about a state in which H is true and she wins her bet; that would be to act wishfully in hope of changing the laws. (At the very least, it would be to act in a way that can only yield good results given different laws—which is wishful if anything is.) Since H is a proposition about the past that determines that Simone will not bet, and Simone sensibly believes that she cannot change the past, Simone knows that she cannot lawfully win her bet. So even if it is true that Simone *would* win her bet if she took it, the truth of H means she should not take that counterfactual as a reason to bet—that would be wishful. Simone should therefore set aside the possibility of winning a car: from the deliberative perspective, she can either lose $\$10$ or walk away.[134]

So, a counterfactual can be true but irrelevant for the purposes of deliberation when you are certain that making its consequent true given its antecedent would violate the actual laws. Simone's mistake was taking an irrelevant counterfactual to be relevant.[135]

More generally: the fact that doing a *would* bring about o is relevant only insofar as that fact tells us that a is a good means of bringing about the end o. That is just to say that we care about acts instrumentally. And if you are certain that facts outside your influence determine that the actual

[134] Of course, you might think that two-boxing in Newcomb's Paradox is similarly wishful since you know that it is very unlikely that you will two-box and win a million (or one-box and not win a million). But the causalist can reasonably deny this: there is nothing incoherent in thinking that you beat the predictor, so you should take seriously outcomes that involve you doing other than the predictor predicted. Contrast this with Simone who knows that she cannot lawfully win the car, so should set aside outcomes that involve her doing other than the laws determine. The thought that we are subject to the laws of nature does not entail that we must reason in a backtracking way when mere act-state *statistical* correlations are involved, as in Newcomb's Paradox.

[135] This does not involve any strong claims about what Simone is *able* to do. Many compatibilists will say that Simone is able to bet despite being determined not to (e.g. Lewis (1981b) and List (2014); more generally, anyone who thinks Simone's abilities are grounded in certain modal facts, such as facts about her dispositions or what she would do if she chose differently, will likely say that Simone is able to bet). I am happy to accept that. Regardless of what she is able to do, I think that Simone should give no deliberative weight to outcomes that would certainly violate the actual laws. Even if there are senses in which Simone can act differently, and those senses are useful in analysing, say, when Simone is morally responsible, I only insist that those facts should not lead to the absurd verdict that Simone should bet. This is to distinguish between Simone's being able to act in a way that would *involve* a law-violation and its being rational for Simone to act *in order* to violate the laws. Thanks to Daniel Nolan for helpful comments here.

world is *not* an $(a \wedge o)$-world, then it is futile to attempt to cause o by doing a. The mere fact that some nearby worlds are $(a \wedge o)$-worlds then provides you with no reason for or against doing a, which means that the truth of the counterfactual $a \rightarrow o$ provides you with no reason for or against doing a. Certain facts about the actual world trump facts about merely possible worlds, so not every (act-outcome, non-backtracking) counterfactual is reason-providing.

Nothing here implies that there is anything wrong with counterfactual deliberation *per se*. You can ask what would happen if you bet on a roulette wheel without worrying about what you are determined to bring about. (At least, those with broadly compatibilist commitments should think so.) We do not ordinarily treat the outcome of a roulette wheel as incompatible with any of your options. So, Simone's case helps us identify which non-backtracking counterfactuals guide action; it does not show that no such counterfactuals do so.

What is required is to incorporate the relevant/irrelevant counterfactual distinction into decision theory so that we can give Simone sensible advice. I spell out problem cases and explain formally why CDT goes wrong shortly, but I present the positive proposal first. The relevant/irrelevant counterfactual distinction can be incorporated into decision theory, meaning we can uphold the policy of causally promoting the good without running into problems with determinism.[136]

3.2.1 Selective Causal Decision Theory

I propose a three-step theory: *Selective* Causal Decision Theory. Rather than simply calculating causal expected utility, you first restrict the set of outcomes that are to be given weight in deliberation (Step 1), adjust your credence in the remaining outcomes (Step 2), and then calculate causal expected utility relative to those adjusted credences (Step 3). Steps 1 and 2 achieve what I proposed in the last section and ensure that only relevant counterfactuals play an action-guiding role. So:

[136] Some causalists might think that Simone should simply bet (call these *die-hard* causalists). What can we say to the die-hard? 1) It is a counterintuitive position; insofar as we are engaged in something like a process of reflective equilibrium, the theory below seems to do better at tying together our theoretical commitments and considered judgements. 2) The die-hard must explain why Simone's betting is not wishful. If Simone is not free to break the laws, then in what sense ought she do something that can only yield good results with a law-violation? 3) Even if the die-hard is right, then good news for CDT and my claim that the causalist standard is reasonable! Nonetheless, I think the considerations raised here make the die-hard position look unattractive.

1. Identify act-state combinations not worth taking seriously for the purposes of deliberation (in decision matrices, we 'grey-out' the outcomes associated with such act-state combinations).
2. For act a, let d_a be the disjunction of dependency hypotheses k such that $a \wedge k$ is not worth taking seriously for the purposes of deliberation. Define the renormalized credence for a, denoted $C_{ra}(\cdot)$ as:
$$C_{ra}(k_i) = C(k_i|\neg d_a)$$
3. You may do a just in case there is at least one k such that $a \wedge k$ is worth taking seriously for the purposes of deliberation and a has maximal *renormalized causal expected utility*, denoted U_R, calculated:
$$U_R(a) = \sum_i C_{ra}(k_i) \cdot u(o_{a,k_i})$$

If dependency hypotheses are conjunctions of counterfactuals, Step 2 amounts to disregarding irrelevant counterfactuals. Step 3 is an expected utility calculation. Step 1, however, needs to be spelled out more precisely. Why are some act-state combinations not worth taking seriously? One reason is familiar from Simone's case: you should give no weight to nomologically impossible outcomes (i.e., those that can only be brought about with a law-violation). So, you should give no weight to $a \wedge k$ if you think that a involves a law-violation given the truth of k. Note an important distinction here. Agents who believe in the truth of determinism will typically think that some outcomes can only be brought about with a law-violation, though they will not know *which* outcomes they are. Step 1 changes nothing for those agents, and they should treat all outcomes seriously. Ordinary compatibilist reasoning following, say, Lewis (1981b) applies in ordinary cases. You consider the various outcomes that you might bring about and know that bringing about some of those outcomes would involve a law-violation. Only when you know that a *particular* outcome involves a law-violation should you cease to give that outcome weight.

Note also that an outcome's involving a law-violation is a sufficient condition for giving it no weight, but it may not be necessary. For instance, if you meet an oracle or time-traveller who tells you that you will not survive tomorrow's battle, then it may be rationally required to disregard outcomes where you survive the battle (see Bales 2016 for discussion of these cases). Theological cases involving predetermination are similar (if I know that God has foreordained that I do not win my bet, the fact that I win my bet in nearby worlds provides no reason to bet).[137] There are likely other kinds of case that follow this pattern; the unifying feature of such cases is that

[137] Though Calvinists might have other reasons not to gamble.

privileged information about the actual world trumps information about merely possible worlds. Having said that, I will focus on law-violations for concreteness and simplicity.[138]

SDT generalises CDT. In ordinary decisions in which you take each outcome seriously, $U_R(a) = U_C(a)$ for all a. So SDT and CDT coincide in ordinary cases, including the cases that typically motivate CDT. Causal expected utility matters *insofar as* you give weight to each outcome featured in the expected utility calculation. When that condition fails, SDT yields more sensible verdicts than CDT. Note also that SDT violates Causal State-wise Dominance, but it obeys a restricted version of that principle:

> **Selective Causal State-wise Dominance:** Say that for all acts f and states k, the outcomes $f(k)$ are worth taking seriously for the purposes of practical deliberation. Then if $o_{f,k} \geq o_{g,k}$, then $f \succcurlyeq g$. Moreover, if for some k $f(k) > g(k)$, then $f \succ g$.

So, we have a further restriction on state-wise reasoning: dominance only applies when each state *leaves open* or is *compatible* (in the nomological sense) with each act. Just as we cannot apply dominance reasoning when acts influence states (states must be independent of acts), we cannot apply dominance reasoning when states determine which act you choose (acts must be left open by states).

3.2.2 Three Cases

3.2.2.1 Betting on the Past

Ahmed(2014a,b) provides the following counterexample to CDT:

> *Betting on the Past*: You are choosing between two bets. a_1 pays out $\$10$ if P and costs $\$1$ if $\neg P$. a_2 pays out $\$2$ if P and costs $\$10$ if $\neg P$. P is the proposition that the actual universe at some past time was in state H and the laws are L; you know that $H \wedge L$ determines that you take a_2 and $\neg(H \wedge L)$ determines that you take a_1.

[138] I have not yet addressed what kind of *epistemic* position you must be in to disregard some outcome. Is certainty, confidence, or knowledge of a law-violation required? For simplicity, I first assume that agents are certain about and know which outcomes involve law-violations. Section 11 considers loosening this assumption.

	P	$\neg P$
a_1	10	1
a_2	2	-10

To see that this case is problematic for CDT, note that it is structurally analogous to Newcomb's Paradox, so CDT recommends a_1 for the same reasons it recommends Two-boxing. In particular, the truth of P is causally independent of your choice[139] and specifies act-outcome dependencies. So $\{P, \neg P\}$ are dependency hypotheses. By Causal State-wise Dominance then, $a_1 \succ a_2$. So, CDT recommends a_1, which is incorrect since the only ways in which a_1 can do better than a_2 involve the laws or past being different.

And again, the reason CDT goes wrong here is the very feature that delivers the two-boxing verdict in Newcomb's Paradox. The advantage of CDT in Newcomb's Paradox seems to be that it lets you treat yourself as independent of the predictor—you reason about the various things you could do and hold fixed their prediction, regardless of whether some of those things are unlikely given their prediction. But here that freedom seems to be a disadvantage—you reason about the various things you could do and hold fixed the laws and past, even though some of those things are impossible given those laws and that past. Ahmed (2013) concludes that rather than treating yourself as free from past constraints, whether powerful predictors or laws, you should take seriously what an act indicates about the past.

But SDT shows that we need not go this far. We can treat ourselves as free from correlations, as in Newcomb's Paradox, while still respecting the authority of the laws. To see this, note that SDT correctly recommends a_2. If P is actually true, then you are determined not to take a_1. Even though you know that if P is true and you *were* to act differently, you would win $\$10$, you also know that doing so would involve non-actual laws. So even though $o_{a_1,P}$ and $o_{a_2,\neg P}$ would be brought about *were* you to act differently than determined, it is wishful to do so: $o_{a_1,P}$ and $o_{a_2,\neg P}$ involve a law-violation and so are not worth taking seriously. SDT therefore calculates:

[139] To see this, note that on any counterfactual semantics that solves Betting on Laws and Betting on History, both the laws and the past are independent of your choice. So any plausible semantics for counterfactuals by the causalist's lights should say that the truth of P is independent of what you do. Ahmed (2014a, pp. 669-71) goes a slightly different route. He takes P refers to the truth of the laws at the actual world—P is true at a world if and only if it is an H-world and L is true at @. So, were you to do otherwise the laws would be different, but the truth of P would remain unchanged. Either way, the truth of P is unaffected by what you do.

$$U_R(a_1) = C(\neg P|\neg P) \cdot 1 = 1$$

$$U_R(a_2) = C(P|\neg\neg P) \cdot 2 = 2$$

And SDT says you should take a_2, which is the correct verdict. So, we can treat ourselves as acting independently of mere statistical correlations without taking law-violating outcomes seriously. We are not constrained by statistical correlations in the same way we are constrained by the laws.

Betting on the Past furthermore highlights the flexibility of SDT. SDT allows you to ignore specific *outcomes* without giving credence 0 to the dependency hypothesis associated with that outcome. CDT, however, can only give 0 weight to an outcome if the associated dependency hypothesis gets credence 0 (standard CDT can only grey-out entire *columns* in a decision table). It is this flexibility that allows SDT to break the parity between Newcomb's Paradox and Betting on the Past. CDT was forced to treat those two cases equivalently. But SDT distinguishes between them: the counterfactual 'if I were to Two-box, I would do better' is true *and relevant*, while the counterfactual 'if I were to bet on *P*, I would do better' is true *but irrelevant.* SDT leverages this distinction to provide sensible advice.

3.2.3.2 *Simone*

Let's return to Simone:

> *Simone*: Simone is offered a bet, which she can Accept or Refuse. If she accepts, she wins a car (valued at $\$10{,}000$) if H is true but $-\$10$ if H is false. Simone knows that H specifies a past state of the world that determines that she will not bet.[140]

	H	$\neg H$
Accept	▨ $10{,}000$	-10
Refuse	0	▨ 0

SDT recommends that Simone refuses:

[140] Again, note that H specifies act-outcome dependencies and is not causally affected by acts, so $\{H, \neg H\}$ are dependency hypotheses.

$$U_R(\text{Accept}) = C(\neg H|\neg H) \cdot -10 = -10$$

$$U_R(\text{Refuse}) = C(H|\neg\neg H) \cdot 0 = 0$$

This contrasts with standard CDT, which says that Simone's high credence in H means that she should accept.[141]

3.2.3.3 Uncertain of Determinism

SDT does not presuppose that you believe in determinism. Consider:

> *Uncertain of Determinism*: You must choose between two bets. The dependency hypotheses are P (which determines that you take a_2), Q (which determines that you take a_1), and R (which does not determine anything). The payoffs are:

	P	Q	R
Accept	10	1	10
Refuse	2	−10	−10

SDT says:

$$U_R(\text{Accept}) = C(Q|\neg P) \cdot 1 + C(R|\neg P) \cdot 10$$

$$U_R(\text{Refuse}) = C(P|\neg Q) \cdot 2 + C(R|\neg Q) \cdot -10$$

This illustrates another advantage of SDT: its advice here is sensitive to your beliefs about P and Q. CDT, however, will recommend a_1 *regardless* of your credences in P and Q. But that is wrong. For example, if $C(P|\neg Q) \approx 1$, then recommending a_1 in an attempt to bring about the impossible $a_1 \wedge P$ is misguided. SDT gives advice that is appropriately sensitive to your credences in these more complicated deterministic scenarios.

[141] A complication: Simone's choice provides evidence about H, so she faces a case of decision instability (if she accepts, the expected utility of accepting is lower than refusing; if she refuses, the expected utility of refusing is lower than accepting). Joyce's (2012) deliberational CDT recommends that Simone engage in a deliberative process, resulting in indifference between accepting and refusing. Others (e.g. Harper 1986) recommend that Simone adopts a mixed strategy that gives non-zero probability to both accepting and refusing. Those responses to instability do not substantially affect the basic point: since it is *impermissible* to accept, being indifferent between accepting and refusing is incorrect, as is acting with non-zero probability of accepting.

3.2.4 The 'No Decision' Response

Insofar as they have addressed them, causalists have typically argued that deterministic cases are not genuine decisions (e.g., Sobel 1988d: 6-10; Joyce 2016). Proponents of this response argue that there is something incoherent in applying decision theory to deterministic cases. Sobel for example argues:[142]

> '[T]he agent . . . cannot consistently so much as even think that both actions are open . . . He must, if consistent think that only one is (though, if he is consistent, he cannot be sure which one that is, unless and until he is sure what he is going to do).' (Sobel 1988d, pp. 6-7)

If this is right, then you cannot think that both options are open in Betting on the Past, so it is a non-decision. CDT therefore does not give the wrong recommendation in Betting on the Past since decision theory only applies in genuine *decisions*. With no decision to be faced, there are no facts about what you ought to do. This 'No Decision' response vindicates CDT, provided deterministic cases really do fall outside the purview of decision theory.

The first thing to say is that, as stated, this argument relies on a strong form of incompatibilism. Sobel takes your choice's being predetermined to imply that you cannot think that more than a single option is open. But if that reasoning holds in Betting on the Past, then it seems to hold *whenever* we reason and suppose that determinism is true. Those convinced of determinism never think that more than a single option is compatible with the past and laws of nature. Requiring multiple open options, in Sobel's sense, threatens to undermine basic compatibilism. Indeed, Sobel says that treating yourself as free in infallible-predictor Newcomb-cases involves a contradiction: on the one hand, by treating yourself as free you commit yourself to the idea that 'a false prediction on the predictor's part [is] itself entertainable, and at least in this sense a possibility' while maintaining 'that a false prediction on the predictor's part is *not* a possibility' (1988d, p. 6). But if determinism is true, then there is always, in a sense, an infallible predictor in the background, so it looks like agents face a contradiction whenever they treat themselves as free.

Indeed, the compatibilist should simply reject Sobel's reasoning. Sobel claims that in deterministic cases you 'cannot consistently so much as even think that both actions are open'. But denying such claims is stock-in-trade for many compatibilists! As already mentioned, we

[142] Though Sobel discusses predictors incapable of error, his comments apply equally to deterministic cases.

should allow that there is a real sense in which you can do multiple things even if, in fact, there is one thing you are determined to do.[143] If we already think that determinism is compatible with your having multiple open options, it is unclear why they should deny that claim in Betting on the Past (certainly both options are *epistemically* open, and are the kind of things we can deliberate about). Conversely, if the causalist wants to utilise Sobel's response, then they must deny that you have multiple open options when you are determined to do one thing. Surely that gets the direction of argument wrong. SDT, however, requires no substantive metaphysical commitments in order to get the right verdict in Newcomb's Paradox.

Joyce defends the No Decision response, though he does so in a way that avoids implicit incompatibilism. Joyce makes a similar claim to Sobel:

> 'If Alice faces [a deterministic case] then, whatever she does, she could not have done otherwise, and perforce, could not have done better. So, [deterministic cases are] a wash when it comes to questions about what Alice should do.' (Joyce 2016, p. 226)

But Joyce frames things more carefully. He also states that 'an agent who deliberates about a decision *which is framed* so that each state *entails* a single act (and outcome) is engaging in an epistemic exercise, not an agential one' (2016, p. 226; first emphasis mine). The thought is that agents face free choices relative to the framing of a decision situation. Joyce can maintain that in most situations you should not be using states rich enough to determine which act you take. For example, when deciding whether to take an umbrella out, you *could* calculate expected utility relative to states like 'Rain and the laws determine you take an umbrella', but you would not face a genuine choice *relative to that framing of the decision*. On the other hand, you would face a genuine decision relative to coarser partitions (like {Rain, No Rain}). Of course, you still do not face a genuine decision in Betting on the Past (the value of outcomes there depends on the truth of deterministic hypotheses, so the states must include deterministic information). In this way, Joyce can treat Betting on the Past as a non-decision without collapsing every choice into a non-decision.

There are good reasons to prefer SDT to Joyce's sophisticated No Decision response.

Firstly, even the sophisticated response is incompatible with a kind of compatibilism that I find independently plausible. Joyce is a compatibilist in the sense that determinism is compatible with genuine decisions, *provided that the value of outcomes does not depend on the truth of propositions that determine your choice*. But why would your caring about the truth of a deterministic hypothesis

[143] Well, at least a common or garden variety kind of compatibilist who broadly follows Lewis (1981b). Other kinds of compatibilist may deny this claim. Thanks to Toby Solomon for discussion here.

affect whether you face a decision? When the compatibilist treats themselves as facing decisions in ordinary cases, they foreclose the possibility of making a choice but subsequently pleading 'well, I didn't really *choose* that—the universe chose, and I merely discovered what the universe decided'. Why should you then adopt that epistemic stance in cases like Betting on the Past? Though you are determined to take a_1 or a_2, you do not know which, so you can coherently treat both options as open and the objects of deliberation. It seems plausible, and in the spirit of compatibilism, that you can take this agential stance while caring about the truth of deterministic theses. SDT allows for this, which is an advantage.

Secondly, Joyce's response involves denying what Ahmed calls 'Soft Determinism' (2014a, pp. 667-669). This is the view that determinism is true and that decision theory should tell rational agents what to do.

> The point of decision theory is to apply to the 'decisions' that you . . . actually face, whether or not those 'decisions' should prove on further investigation to have been free in the incompatibilist's sense. (Ahmed 2014a, pp. 667-669)

I agree with Ahmed: there are better and worse courses of action in deterministic cases, and we want decision theory to rank better courses of action over worse ones. More complicated cases highlight this problem for the No Decision response:

> *Betting on the Past II*: You are choosing between two bets. The dependency hypotheses are P (which determines a_2), Q (which determines a_1) and R (which determines a_1). The payoffs are:

	P	Q	R
b_1	0	-100	200
b_2	100	0	0

The No Decision response treats this as a non-decision. But surely we can distinguish between the advice we would give Barbara (who is practically certain of Q) and Ian (who is practically certain of R). If we said to both Ian and Barbara, 'do what you will, the correct decision theory can say no more', they would rightly point out that such advice is insensitive to their disagreement about the world. There is an asymmetry in their perspectives, which results in an asymmetry in what the best means to each agent's ends are. Barbara does better by her lights in taking b_2, as Ian does better by his lights in taking a_1. It is true that there are outcomes that agents should set aside as they deliberate, but that does not mean we cannot provide them with guidance.

Finally, the No Decision response gets things wrong in the following:

> *Partial Determination*: You must choose between c_1, c_2, and c_3. The dependency hypotheses are P (which determines $\neg c_1$), Q (which determines $\neg c_2$), and R (which determines $\neg c_3$). The payoffs are:

	P	Q	R
c_1	100	0	0
c_2	10	0	10
c_3	10	10	0

SDT says: $U_R(c_2) = U_R(c_3) = 10$, while $U_R(c_1) = 0$. SDT says that both c_2 and c_3 are permissible, which is the correct verdict.

Saying that this is a non-decision is incorrect. Even by Joyce's and Sobel's lights, each state allows for a genuine choice. It would therefore be incorrect to say that the whole decision is a wash. But it would also be misguided to give weight to every outcome represented—that would lead to taking c_1 (given sufficiently high credence in P). And taking c_1 is absurd, no matter how confident you are in P.[144] The only sensible policy is to calculate expected utility based on just the non-greyed-out outcomes. The No Decision response is too coarse-grained to handle intermediate cases like Partial Determination.

One possible move on behalf of the No Decision response is to say that while you do face a choice in Partial Determination, you only find out which choice that is after learning which

[144] Though this case again involves instability, we can still cause problems for deliberative versions of CDT. Say you have initial credences: $C(P) = C(Q) = C(R) = \frac{1}{3}$, $C(P|c_1) = C(Q|c_2) = C(R|c_3) = 0$, and $C(Q|c_1) = C(R|c_1) = C(P|c_2) = C(R|c_2) = C(P|c_3) = C(Q|c_3) = \frac{1}{2}$ (i.e. each act makes certain that the incompatible dependency hypothesis does not hold and leaves the remaining two hypotheses equally likely). Then, using Joyce's (2012) dynamics, you get the equilibrium: $U(c_1) = U(c_2) = U(c_3) = \frac{100}{9}$, with $C(P) = \frac{1}{19}$, $C(Q) = C(R) = \frac{9}{19}$, $C(c_1) = \frac{17}{19}$ and $C(c_2) = C(c_3) = \frac{1}{19}$. This means that c_1 is permissible, and moreover that you ought to be confident that you will perform c_1.

dependency hypothesis is true. For example, if P is true, then you faced a choice between c_2 and c_3 all along. But this yields no more action-guiding advice than just calling the case a non-decision. After all, you cannot tell *which* options you face a choice between while deliberating. But then you cannot know which outcomes to exclude from deliberation, leaving us with the initial dilemma: either provide no advice or provide advice based on all outcomes. Neither option is satisfactory.

Though I have rejected the No Decision response, it is worth highlighting an important methodological agreement my view and Joyce's version of that response: we both think that the framing of a decision problem matters. Our disagreement concerns *how* it matters. While Joyce thinks that building deterministic information into state-descriptions changes *whether* you face a decision, I think that it changes the *nature* of your decision (by affecting which outcomes are worth taking seriously). So, in cases like Simone's, where deterministic information must be built into state descriptions, the correct response is to restrict the set of outcomes worth taking seriously, not to give up on agency altogether. Nonetheless, I agree that the carving up of states plays an important role in handling deterministic cases, which is reflected in the fact that SDT calculates expected utility relative to dependency hypotheses.

3.2.5 Avoiding Evidentialism

Next, I want to consider whether SDT deserves to be called a *causal* decision theory. You might think that in Step 1, SDT covertly appeals to the evidence an act provides to evaluate that act. That is, we grey-out $o_{a,k}$ because a provides evidence against k (albeit evidence of the particularly strong kind that k is impossible). This would count as evaluating acts based on their news-value. But the hallmark of causalism is that acts are not evaluated based on their news-value; when news-value and causal efficacy diverge, it is causal efficacy that determines what you should do. So, if SDT is taking account of news-value, then it is an *ad hoc* compromise.

It is important to stress the rationale behind Step 1 of SDT. You do not grey-out $o_{a,k}$ because of your conditional credence $C(o_{a,k}|A) = 0$. That would count as *ad hoc* evidentialism. Instead, SDT greys-out outcomes because they are futile to attempt to bring about—doing so would involve a law-violation. Certain facts about the actual world play a *structuring* role in the SDT'ers deliberation. To take the possibility of determinism seriously is to give up on being free to break the laws, which means that your beliefs about what is nomologically possible form fixed points in your deliberation. Of course, when you know what the laws determine, your acts may have news-value as well. But that does not mean we disregard outcomes *because* of that news-value. In Betting on the Past, for example, you disregard the possibility of winning $\$10$ because you respect the laws: knowing that you are determined not to win $\$10$, you structure your deliberation around that fixed point. Step 1 of SDT does not appeal to your conditional views; rather, it appeals to your unconditional views about what is nomologically possible. And that is importantly different to caring about news-value.

This defuses a potential objection from Ahmed. While discussing Betting on the Past, Ahmed (2014a, pp. 678-9) claims that any theory deserving to be called 'causal' is forced to give some weight to greyed-out outcomes. If this is correct, SDT does not count as causal. Why does Ahmed think that the causalist is committed to giving non-zero weight to greyed-out outcomes? Because the hallmark of CDT is the use of counterfactual reasoning, and counterfactual reasoning involves thinking about possible worlds that certainly differ from our own. In Betting on the Past then, the causalist is supposedly forced to ask 'given P, what would the world be like if I took a_1?', even though they know that P makes the actual world one in which they do not take a_1. Ahmed concludes:

> 'CDT regards worlds that are open to a free agent as those that would obtain were she to act otherwise than she actually does, even if those worlds are certainly non-actual.' (Ahmed 2014a, p. 679)

Ahmed is right that *unadorned* CDT regards certainly non-actual worlds as open to free agents, but SDT makes no such requirement. Ahmed regards the kind of counterfactual just described as definitive of causalism. I think that a more nuanced position is required: counterfactual thinking is definitive of causalism *at the level of expected utility calculations*. The core of causalism is that expected utility is a matter of expected causal efficacy. But before calculating expected utility, the causalist is entitled to use non-counterfactual reasoning to restrict the set of outcomes that go into the expected utility calculation. The actual world matters, even for the causalist! So SDT'ers care about causation, though they deny that all outcomes are worth causally promoting.[145]

[145] Ahmed (2015) raises a distinct worry for views like SDT. He argues that every Newcomb case can be viewed as a weighted lottery between a certainly correct predictor and a certainly incorrect predictor. He argues that views like SDT will then recommend one-boxing in *every* Newcomb case (since when framed as a weighted lottery between certainly correct and incorrect predictors, we will have to grey-out outcomes such that SDT agrees with Evidential Decision Theory). This does not seem to be a problem for SDT, since Ahmed's argument relies on calculating expected utility relative to the partition: {Certainly Correct Predictor, Certainly Incorrect Predictor}. But these states are not causal dependency hypotheses, so we cannot use them in SDT's expected utility calculation. This again highlights the importance of framing: when determinism is involved, not just any states will do.

It is important not to miss the wood for the trees here: in ordinary cases, SDT agrees with CDT, and SDT agrees with CDT in motivating cases like Newcomb's Paradox. In cases where SDT builds on CDT, it does so in a way that respects the causalist intuition that news-value is irrelevant to decision-making. SDT still deserves to be called causal.

But you might still worry that SDT opens the door to evidentialist reasoning, even if it does not itself rely on evidentialist reasoning. In particular, you might worry about an objection raised by Seidenfeld (1984; see also Sobel 1988 and Ahmed 2015). Seidenfeld objects to theories that treat Newcomb cases in which $C(\text{Predictor Correct}) = 1$ differently to those in which $C(\text{Predictor Correct}) = 1 - \epsilon$ for any $\epsilon > 0$, since he thinks that such an ϵ-decrease cannot make a difference to what you ought to do. Now this objection does not directly target SDT, since it greys-out based on your views about what is nomologically possible, not merely your confidence in the predictor's accuracy. But a similar objection might arise: SDT distinguishes between cases in which you are *certain* that some outcome involves a law-violation and cases in which you are merely *confident*. This raises a question that I have hitherto ignored: what kind of epistemic situation must you be in to disregard some outcome? I now turn to that point.

3.2.6 What is Futile?

You should disregard an outcome when it is not worth taking seriously for the purposes of practical deliberation—when it is futile to attempt to bring it about. But when are you entitled to do this?

The easiest cases are those in which you are rationally certain that some outcome involves a law violation. If you are rationally certain that you cannot lawfully bring some outcome about, then your deliberation should be structured around that fact.

But what about cases involving less than complete certainty? Consider the following:

> *Betting on the Past III*: You are choosing between two bets on P: a'_1 and a'_2. You are confident but not certain (your credence is $.99$) that P determines that you will take a'_2, and you are confident but not certain (your credence is $.99$) that $\neg P$ determines that you will take a'_1.[146]

[146] There are various ways this could be: you could be $.99$ certain that some deterministic theory holds, or you could be certain that some system of laws holds that is deterministic apart from occasional indeterministic blips. My proposed solutions treat these versions of the case the same. There is a related case: the one in which you know that

	P	$\neg P$
a_1	10	1
a_2	2	-10

Given that you are not certain that any outcome involves a law-violation, should you treat this case like Betting on the Past or Newcomb's Paradox? That is an interesting question, and I remain broadly neutral, partly because I have no firm intuitions about this case, and partly because a full argument for either position would be beyond the scope of this chapter. Instead, I sketch three ways of analysing futility. SDT can incorporate any of these analyses and so deliver different verdicts in Betting on the Past III.

Firstly, we could adopt a *strict* analysis of futility. On this account, for an outcome to not be worth taking seriously, you must have credence 1 that it involves a law-violation. This means treating Betting on the Past III like an ordinary Newcomb case: every outcome is worth taking seriously, which means you should take a'_1.

Some might worry that the strict analysis is too strict. After all, it is extraordinarily *unlikely* that your doing better by taking a'_1 is compatible with the laws. Does this mean that you ought not take a_1'?

I am not sure. Insofar as there is an intuition that you ought not take a'_1, it is unclear how much weight to put on that intuition. And the proponent of the strict analysis can point to a difference between Betting on the Past and Betting on the Past 3: your winning $\$10$ without violating the laws is possible in the former but not the latter. (This explains why an ϵ-decrease in credence might be significant, *contra* Seidenfeld. A shift from $C(\neg X) = 1$ to $C(\neg X) = 1 - \epsilon$ can signal a shift from X's being impossible to possible.) True, it is unlikely that P is true and you take a'_1 in Betting on the Past III, but there is nothing incoherent involved in taking a'_1 and winning $\$10$ (it is just unlikely). And causalists already think that we need to take unlikely act-state combinations into account when deciding between options (like Two-boxing when the predictor guessed One-box). In Betting on the Past, you are certain that you cannot lawfully win $\$10$, so the fact that you do not win $\$10$ should act as a fixed point as you structure your deliberation.

P determines that the chance of a_2 is $.99$. I will not settle what you should do in that case since it is a case of thoroughgoingly *indeterministic* laws; such cases certainly raise challenges, though as they are not the focus here, they must be for future work.

But you are not certain of that fact in Betting on the Past III, so it might make sense to treat winning $\$10$ as a live possibility. Causalists should not be misled by the fact that taking a'_1 provides strong evidence against P; that is just the kind of news that the causalist sets aside in Newcomb's Paradox, and they should set it aside here.

A second approach would be to adopt a *threshold* analysis of futility. For an outcome to not be worth taking seriously, we might insist only that you have high credence that it involves a law-violation.[147]

The threshold view faces the standard worries that thresholds are seemingly sharp and arbitrary. Say that we set the threshold at $.99$. Then we might ask what is the real difference between $C(\text{Law Violation}) = .99$ and $C(\text{Law Violation}) = .9899$? How could that miniscule decrease in confidence affect whether you take some option seriously? And why pick $.99$ in the first place?

At this point, the defender of the threshold analysis can make use of the moves that get made in response to Sorites sequences. We could say that there is a threshold, though we may not be able to work out where it is. Or we could say that there is a vague threshold. This strikes me as plausible: we can point to paradigmatic cases where some outcomes are futile (Betting on the Past), and we can point to paradigmatic cases where no outcomes are futile (Newcomb's Paradox). Between those cases, there might be a range of indeterminate cases. If you are $.98$ confident that your winning big by Two-boxing involves a law-violation, then perhaps it is indeterminate whether you ought to Two-box. Freedom is a tricky concept, and it seems plausible that we might sometimes be neither determinately free nor unfree to bring some outcomes about. Clearly, more needs to be said here. I simply want to point out that defenders of the threshold view will be able to draw on the tools developed elsewhere to help deal with problematic threshold concepts.

A final strategy would be a *knowledge-based* analysis of futility: you should treat some outcome as not worth taking seriously when you *know* that it involves a law-violation.[148] On this view, it is not partial belief but knowledge that determines the fixed points around which deliberation should be structured.

[147] I take Joyce (2016) to advocate a kind of threshold account, though his analysis of when to disgregard outcomes differs slightly from SDT's.

[148] This suggests related approaches: belief approaches, justified belief approaches, etc. Hopefully it is clear how other concepts could be substituted into this analysis.

Some might be uncomfortable introducing a concept like knowledge into decision theory. But Weatherson (2012) argues that decision theorists cannot ignore knowledge. On Weatherson's view (2012, p. 77), you know *p* if and only if it is legitimate to write *p* as an outcome in your decision table (similarly, you know that a state does not obtain if and only if you can legitimately leave that state off the decision table). So, knowledge may play a role in decision theory: given that agents like us are rarely certain about things, knowledge helps us to understand what goes into our decision tables in the first place. Now, Weatherson does not talk about the kinds of cases under consideration here. But it seems natural to extend his account to supplement SDT; indeed, Weatherson is concerned with how decision problems should be structured, and I am arguing that one structuring principle is that you should disregard law-violating outcomes. So, we might suggest: an agent can legitimately grey-out an outcome if they know that the outcome would involve a violation of the actual laws. I will not try to give a further analysis of knowledge here, but note that this approach may take into account whether your beliefs are justified, how you came to have your beliefs, the stakes of the case, and so on.

The strict analysis will say that you ought only disregard outcomes in Betting on the Past. The threshold analysis will say that you should disregard outcomes in both Betting on the Past and Betting on the Past III (given a choice of threshold below $.99$). The knowledge-based analysis will say that it depends on what you know in each case. I have argued that each of these strategies is plausible, though it is beyond the scope of this chapter to argue that any one strategy is best. What matters is that whichever option is taken, SDT can be supplemented with a plausible account of when some outcomes are not worth taking seriously.

3.3 Conclusion: Causal Compatibilism

Determinism raises serious questions about the nature of rationality. I have argued that we can respect Causal State-wise Dominance *without* betting against our credences in Ahmed's Betting on the Laws. If we acknowledge the tension between determinism and the ability to do otherwise, we need something like impossible worlds to capture the modal robustness of the past and the laws. And once we secure that modal robustness, CDT delivers plausible verdicts in Betting on the Laws.

The lesson from Betting on the Past is that we must refine Causal State-wise Dominance. Not only must states be act-independent to apply dominance reasoning, those states must not determine what you do. I have argued that we still face genuine decisions when states determine—we can deliberate even if our acts might be predetermined. SDT allows us to deliberate when our acts might be predetermined, and it respects an appropriately modified dominance principle. This allows SDT to carve between Newcomb's Paradox and Betting on the Past, agreeing with CDT in the former but departing and giving more plausible verdicts in the latter.[149]

So, deterministic cases do not break the stalemate between causalist and evidentialist. They do force us to clarify and refine the causalist view, but they do not undermine the basic motivation behind causalism.

[149] Note that there is a class of related challenges that I have not addressed, those involving foreknowledge and chancy processes (see Rabinowicz 2009, Price 2012, and Bales 2016). Future work is to show how SDT interacts with those challenges.

Chapter 4

Beyond Stalemate: Evidence, Action, and the Role of Decision Theory

4.0 Stalemate, or Standoff?

Let us grant that the causalist-evidentialist debate is at a stalemate. Where should we go from here? Horgan (2017, p. 41) says that acknowledging the stalemate is liberating 'to the extent that one now feels justified in not worrying about the opposition and in just going ahead and constructing a decision theory that consistently yields the answers which one has already decided are the right ones'. And indeed, there is something liberating in beginning with the judgements you take to be right, regardless of whether you can convince someone else of the correctness of those judgements. We each look for a theory that rationalises what we already take to be reasonable—either one- or two-boxing as the case may be.[150]

Of course, we might wonder whether 'stalemate' is quite the right word for the situation I have just described. In a stalemate, *both* parties acknowledge that there is a draw. But in the situation Horgan has described, each party sticks to their guns and refuses to back down in light of criticism from the other. So, I suggest that *standoff* is a better way of thinking about the debate between causalist and evidentialist. Each party has an internally consistent position and cannot be persuaded to back down from that position.

How then should we each go about 'constructing a decision theory' that seems right to us? Horgan says that we construct a theory around the verdicts we already take to be the right ones. But I think we can say more. We might be after more than reasonable verdicts in *individual cases* as we construct a theory. We might ask what properties we want our decision theory to have—what we think a theory of individual rationality should look like—and construct a theory with those properties. Following Horgan, I refer to these as *meta-considerations*—facts not about what is rational, but facts that inform our decision to adopt one standard of rationality over another.

[150] Bales (2018b, pp. 807-809) similarly defends pluralism as a response to Newcomb's Paradox. On Bales' view, there are multiple sharpenings of the word 'rational' (causalist-rational and evidentialist-rational) and so, on a supervaluationist approach to vagueness, both one- and two-boxing are indeterminately permissible. I take Bales' position to be compatible with Horgan's. Even if what is rational *simpliciter* is indeterminate (following Bales), each *individual* might construct (or adopt) a decision theory rationality that best fits their intuitions (following Horgan). So, we must distinguish 'rational pluralism' (that there are multiple, legitimate standards of rationality and you are free to adopt any of them) from what might be called 'indeterminacy monism' (that there is an overarching perspective of rationality that requires you to judge one-boxing to be indeterminately permissible). I adopt rational pluralism here.

What is a meta-consideration?[151] Consider the term 'bald' and say that you are tasked with constructing a theory of baldness. You accept that 'bald' is a vague term, so there is no precise number of hairs that separates the bald from the non-bald. You might (in the spirit of Horgan) ask which verdicts about baldness *you* already accept and construct your theory around those verdicts. But another natural question to ask is what you want a theory of baldness *for*. Perhaps you want a theory of baldness to help identify candidates for a Patrick Stewart lookalike contest—if so, then some theories of baldness stand out as appropriate. Or perhaps you want a theory of baldness to help identify potential wig-buyers—if so, then different theories of baldness will stand out as appropriate. In this way, we go about constructing our baldness-theory based not just on verdicts about cases, but based on what it takes for a theory to be fit for purpose. So, here I focus on meta-considerations of the form 'Because we want our theory to do X, a theory with property Y is more fit for purpose than one that lacks property Y'.

What then do we want a decision theory for? Well, *I* want a decision theory to provide action-guiding advice—the kind of advice that I can understand, interpret, and implement to steer my behaviour. I do not have a particularly precise characterisation of what it takes for a theory to be action-guiding. Instead, I will be to point to two features of CDT that strike me as paradigmatic *failures* to provide action-guiding advice. So, though I might not be able to convince the committed causalist to abandon CDT for EDT, I end up as an evidentialist when I go about constructing a view with the properties that characterise an action-guiding theory.

Now, action-guiding advice is not the only thing you might want a decision theory for. But it a natural desideratum to focus on in the context of the debate between causalist and evidentialist. A different thing you might want a decision theory for is to help us analyse or explicate some notion of value. But plenty of causalists accept that U_E explicates a coherent notion of value (see Hutteger and Rothfus Forthcoming, Section 1)—what causalists deny is that you should *act* on the basis of value as explicated by U_E. Or you might want a decision theory to predict and explain behaviour. But the causalist will accept that plenty of people one-box, and the story about how probability and causation interact in our reasoning likely complex—U_E and U_C may each play a partial predictive-explanatory role. Or you might want a decision theory for evaluative purposes. But if causalism and evidentialism are both internally consistent standards of rationality, then it is hard to see what force a judgement of causal-irrationality has if someone adopts the evidentialist standard (and vice versa). When it comes to evaluating somebody, the

[151] The meta-considerations I consider here are different to those that Horgan (1981) considers. Horgan is largely concerned with *pragmatic* meta-considerations (facts about which standard of rationality results in agents doing best), but I widen the net and consider any fact that speaks in favour of adopting some theory over another.

only kind of evaluation *they* will care about is the kind made relative to a standard they already accept. So, when looking for reasons to adopt either CDT or EDT, it seems natural to focus on the action-guiding side of decision theory. That is where the differences are sharpest and where we might expect those differences to speak most clearly in favour of one theory over the other.

EDT does better as an action-guiding theory on two counts.

4.1 Decisions and Intentions: You've Got to Be Able to Do What You've Got to Do

Claim: if you advise me to do something that I cannot bring myself to do, then your advice is not action-guiding. That is to say that action-guiding advice must be the kind of thing that I can implement.

CDT can require that you do things you cannot bring yourself to do. Moreover, the *reason* you cannot bring yourself to do those things is because you consistently follow CDT. In virtue of following CDT then, you cannot implement CDT's advice. Consider the following:

> *Flexi-Bus*: As in Death in Damascus, you must decide between Aleppo and Damascus. And again, Death is an excellent predictor of your decisions (Death gets things right 90% of the time). There are two asymmetries here: (i) There is a sweetener if you meet Death in Damascus, and (ii) a friend claims to have seen Death in Aleppo this morning, so you are 99% confident that Death is in Aleppo. If you go to Damascus, you cannot change your mind. If you got to Aleppo, you travel by FlexiBus, which allows you at any stage on the trip to change your mind and go to Damascus. (Note that the trip is a long one, so there are many opportunities to press the button.)

	Death in Damascus	Death in Aleppo
Go to Aleppo	10	0
Go to Damascus	$0 + s$	$10 + s$

The two important things to note here are: (i) initially, CDT recommends going to Aleppo, and (ii) if you become confident that you will go to Aleppo, CDT recommends going to Damascus. So, if you try to implement CDT's advice to Go to Aleppo, you do not believe you will make it to Aleppo. Indeed, given the sweetener in Damascus, if you try to go to Aleppo you are *more* likely to end up in Damascus than Aleppo.[152] So, CDT recommends one thing (go to Aleppo), but in virtue of following CDT's advice you are confident that you will end up doing something else.

In what sense then is CDT providing action-guiding advice when it recommends going to Aleppo? 'Go to Aleppo' is something we would ordinarily say that you can do (there is no forcefield preventing you from boarding and staying on the bus), that CDT recommends, but that you cannot bring yourself to do *because* you follow CDT. This advice is self-undermining.

I suggest the following constraint on what it takes for a theory to be action-guiding:

> **Action-Implementation:** If in some decision situation, theory T says that the optimal act at some time is ϕ, and in virtue of following T you do not believe that you will ϕ, then T's advice is not action-guiding.

Crucial here is that a theory's advice is not action-guiding if you cannot implement the optimal act *in virtue* of following that theory. In Flexibus, what makes it the case that you cannot go to Aleppo is that you consistently follow CDT's advice. It is not your own weakness of will or similar that prevents you from getting to Aleppo (in that case, the action-guiding failure would be with you, not the theory). The reason that you cannot do what CDT recommends is that you follow CDT. So, CDT's advice in FlexiBus is no help in navigating the world—it identifies things that I would like to be able to do, but the structure of the theory itself prevents me from doing those things.

EDT is action-guiding. Say that EDT at some time recommends ϕ (i.e., $U_E(\phi)$ is maximal). Let $d\phi$ be a decision to perform ϕ (which might include the initial stages of performing ϕ) and let us restrict ourselves to the class of reasonable decisions in which for each act ψ and outcome o, $C(o|\psi) = C(o|\psi \& d\phi)$.[153] Then by definition for each ψ, $U_E(\psi) = U_E(\psi|d\phi)$. So, on

[152] A word on equilibrium credences here. Say that rather than going to Aleppo in Flexibus, you reason your way to equilibrium credences, as Joyce (2012) recommends. CDT's advice may still not be action-guiding, even if you reason your way to equilibrium credences before acting. Let's tweak the case so that whatever you decide, Go to Aleppo or to Damascus, you can at any time pay a small fee to change your mind, but you can only change your mind once. At equilibrium, you judge both Aleppo and Damascus to be permissible (both are unstable, so get assigned non-zero probability at equilibrium). But whichever you decide, you know that, very likely, at some later stage you pay a small fee to go to the other place—if we include enough opportunities for reconsideration, chances are you will reconsider once. So again, for each option that CDT recommends as choiceworthy at equilibrium, you do not believe you can implement that option in virtue of consistently following CDT.

[153] Strictly speaking, we only need assume that your conditional credences are such that if ϕ is U_E-maximal, then ϕ is still U_E-maximal after conditionalising on your decision to perform ϕ. This holds in most canonical cases in the literature, including cases like Death in Damascus and the Psychopath Button. In cases where this assumption fails, EDT may fail to provide action-guiding advice. But such cases are strange ones—imagine a predictor who places a million in Newcomb's Paradox if they predict that you *decide* to one-box, regardless of whether you actually end up

following EDT's advice—deciding on ϕ when EDT recommends ϕ—your views about what is optimal are unchanged. EDT's advice is stable in cases like FlexiBus and Death in Damascus, while CDT's is not, which means you can implement what EDT, but not what CDT, recommends.

The natural response on behalf of the causalist is to say that in cases like FlexiBus the world is not structured to allow you to pursue a unified course of action. Hare and Hedden (2016, p. 613) entertain the idea that in cases of instability 'a lack of commitment to your decisions is precisely the right attitude to have'. As you make decisions, you receive evidence about where Death is, so you change your views about what is preferable and defect from your initial strategy. Such predictable defection might just be the appropriate response to playing against Death.

That response is of course coherent—by the causalist's lights there is no reason to think that you will make it to Aleppo (if your decision provides evidence that Death is in Aleppo). But we are not playing the *internal* consistency game here. Rather, we are looking for reasons to adopt one standard of rationality over another. CDT says that you should expect to deviate from what you currently think optimal when playing against Death. But the reason you cannot pursue an optimal plan here is not just that you are playing against Death. Rather, what *makes* it infeasible to carry out optimal plans is that you follow CDT's advice. There is a legitimate standard of rationality—the evidentialist one—that allows you to pursue dynamically unified action when playing against uncanny predictors like Death. So, it is not the world itself that prevents us from pursuing dynamically unified action—it is the way the world is *coupled* with our theory choice. When deciding between theories then, we cannot plead 'that's just the way the world is'. EDT satisfies a desideratum for a decision theory that CDT does not (even if the causalist can rationalise away that desideratum from their internal perspective). Insofar as we prefer an action-guiding theory to a non-action-guiding one, we should prefer EDT.

4.1.1 Causalist Responses

one- or two-boxing. In such a case you are rewarded for making a decision then deviating from it (the predictor rewards one-boxers, but even more so rewards people who decide to one-box but end up two-boxing). It is hard to know what *decision* theory should say in cases where you are rewarded for *deviating from a decision*—so I set such them aside here.

I consider two natural moves on behalf of the causalist—one that supplements CDT with some theory of plans (or resolutions, commitments, intentions, etc.) and the other that refines our theory of options.

4.1.1.1 Resolutions, Commitments, Intentions

Firstly, you might acknowledge that *unadorned* CDT fails to provide action-guiding advice in FlexiBus. But you might think that we already need to adorn standard decision theory to coordinate between our time-slices. There are three broad ways of achieving this: resolutions that alter preferences (see McClennen 1988, 1889), commitments that override preferences (see Gauthier 1994), and intentions that coordinate decisions (see Bratman 1987).

My response to each of these is that the more comprehensive the solution, the greater the costs of that solution. And since we are in the game of providing meta-considerations, it is an advantage of EDT that it achieves for free what CDT only achieves with costly or mysterious supplements.

Firstly, consider an agent who makes resolutions that *influence* their future preferences (suggested by McClennen 1988). That is, if at time t I resolve to ϕ at time t^{+}, then at t^{+} my preference is to ϕ.

Plausibly, we do often prefer to carry out past intentions, and sometimes the mere act of making a resolution might influence our underlying tastes (my resolution to order dumplings tonight might generate a desire for dumplings). What is less plausible (indeed, to my mind implausible) is that those preferences are strong enough to decisively influence our behaviour in cases of interest, like FlexiBus.[154] Say that the causalist prefers to carry out resolutions. Now consider that they have resolved to go to Aleppo, that they are on the FlexiBus to Aleppo, and that they are confident they will meet Death in Aleppo. For it to be the case that the causalist can make it to Aleppo, either (i) the causalist values resolve so much that they prefer a high probability of death (and carrying out a resolution) over a high probability of survival, or (ii) there is some psychological mechanism by which their resolve has altered their tastes and they now prefer a high chance of death (by going to Aleppo) over a high chance of survival (by going to Damascus). Both (i) and (ii) are absurd—for our decision theory to be action-guiding, we ought

[154] Sobel (1986, pp. 537-538) raises a similar criticism of McClennen's response to dynamic Allais cases.

not presuppose that agents have such idiosyncratic preferences. So, resolution-influenced preferences do not enable CDT to provide action-guiding advice in all cases.

Secondly, consider an agent who makes commitments that *override* their preferences. On this account, agents have something like a 'basic capacity' to form commitments and carry those commitments out, regardless of their future preferences. That is, if at time t I commit to ϕ at time t^+, then at t^+ I carry out ϕ *even if* I then prefer not to do so (cf. Gauthier 1994).

I personally find a basic capacity for commitment deeply mysterious. This approach posits something over and above preferences that guides action, so it is outside of the scope of decision theory as standardly conceived. Moreover, it is a paradigmatically *non-consequentialist* approach to choice—what I ought to do at some time depends not just on what best promotes consequences at that time, but on what I happened to commit to in the past (I discuss this further in Chapter 6). And even if we accept a non-consequentialist picture of rationality, there is a significant explanatory burden that accompanies a basic capacity account of commitments. I often make commitments but find myself deviating from them for a good reason. In cases of instability, the causalist gets evidence on making a decision that, by their lights, shows that decision to be wrongheaded. The defender of commitments might say that the causalist is objectionably weak-willed if they deviate from their past commitment. But if the causalist thinks that they can escape Death by deviating, I do not blame them for doing so. Requiring that agents have the basic capacity to steadfastly meet a near-certain death is a high bar for what it takes to be rational. So, I am suspicious of positing such a basic capacity. And since EDT achieves dynamic unification without positing mysterious commitments, I prefer EDT.

Finally, consider an agent whose intentions play a *coordinating* role. That is, by intending to ϕ I provisionally structure my deliberation around the fact that I will later ϕ. This view, defended by Bratman (1987), is psychologically plausible. Unfortunately, it fails to address the action-guiding worry for CDT.

Everyone agrees that intentions (of the coordinating kind) had better be revisable. If I intend to cross the road, I might structure my deliberation around my road-crossing. But if I learn that a runaway lorry is hurtling down the road, then I should re-open deliberation and reconsider whether crossing the road is the best means to my ends. Now, the details of *when* intentions are rationally revisable, or what kinds of intention-revision habits are rational, is up for debate. But revisable they must be.

My claim is that any minimally plausible theory of intentions will say: *if* you get strong evidence that you die by carrying out some intention, you are permitted to revise that intention.

Though I cannot discuss every possible theory of intention-revision, we might take the above claim as a stipulative constraint on a reasonable theory. As proof of concept that at least some prominent theories are reasonable (in my sense), I consider Bales' (2020) view, which specifically addresses how CDT interacts with intentions in cases of instability. Stakes-shifts form a crucial part of Bales' theory. In particular, intention-revision is permissible when the stakes *go up*—if I intend to place a bet at the casino and later find out that the chips are for thousands of dollars, not individual dollars as I previously thought, then I should reconsider my intention to bet. So in the Psychopath Button, Bales (2020, pp. 798-799) argues that the causalist will reason:

> *Step 1*: CDT recommends pushing the Psychopath Button, so I form an intention to do so.
>
> *Step 2*: The stakes have gone up (previously, I faced a high probability of some moderate good or the status quo; I now face a high probability of death).
>
> *Step 3*: I re-open deliberation and re-calculate U_C. Now that I think myself likely to be a psychopath, CDT recommends refraining. So, I intend to refrain.
>
> *Step 4*: There is no increase in the stakes, so I do not re-open deliberation. Hence, I carry out my intention to refrain.

This model of deliberation in the Psychopath Button strikes me as reasonable.[155] But now we face the action-guiding challenge afresh. At Step 1, if I am self-aware, I know that I will not carry out my intention to Push (indeed, I will end up doing the opposite, Refrain). And yet CDT recommends that I Push at Step 1. What precisely is the status of CDT's advice at Step 1 then? CDT recommends forming an intention to do something that I believe I will not do, moreover that I will not do in virtue of following CDT. So, I cannot implement CDT's advice in virtue of following CDT.[156] A reasonable theory of intentions leaves the causalist with the action-guiding worry.[157]

[155] You might worry that your initial intention to Push the Psychopath Button is not rationally revisable because you always *knew* that on implementing that intention the stakes would go up. And perhaps you ought not revise intentions in light of such foreseeable stakes-changes. See Bales (2020, pp. 801-802) for a response to this point.

[156] Indeed, on Bratman's (1987) view of intentions it is *incoherent* for me to intend to do something that I believe that I will not do. So again, *because* you follow CDT, the very thing that CDT says is optimal is not something that you can coherently intend to do.

[157] Bales (2020, p. 800) suggests that in cases where intention-formation is costly, the causalist might see that forming an intention to Push is futile and so skip straight to forming an intention to Refrain. But this will not solve

So, I am still unsure about what it means to act on the basis of CDT's ranking of option. We could supplement CDT, but doing so is either implausible, mysterious, or fails to address the action-guiding challenge in full. The EDT'er, however, has stable preferences as they implement EDT's advice. So, I know precisely what it means to act on the basis of EDT's advice.

4.1.1.2 Cohesive Expected Utility

A second strategy from Meacham (2010) involves not *supplementing* CDT but shifting the target of instrumental utility maximisation. Meacham thinks that rather than assessing acts directly, we should assess acts as parts of strategies. He uses the term *comprehensive strategy* (p. 53) to refer to what I have been calling a plan—a specification of what you do at each choice node in a decision tree. (Recall from the Introduction: a decision tree is a directed graph that begins with a root node, in which each subsequent node represents a choice that either you or nature makes, with each sequence of choices terminating in an outcome.) He then says that an act is rational if it is part of an optimal comprehensive strategy.

Denote some such comprehensive strategy CS. Let C_i and u_i be your initial credence and utility functions respectively. The *causal cohesive expected utility* of strategy CS is (Meacham 2010, p. 68):

$$CEU(\mathrm{CS}) = \sum_j C_{\mathrm{i}}\big(\mathrm{CS} \rightarrow o_j\big) \cdot u_{\mathrm{i}}(o_j)$$

Causal CEU-maximisation says that you may perform an act just in case it is part of a CEU-maximising comprehensive strategy. For example, say that at the outset of some decision tree there is one comprehensive strategy that causally promotes the good by the lights of your initial credences and utilities: 1) go to the pub, 2) have a drink, then 3) order a taxi. Even if after going to the pub and having a drink your preference is to cycle home, cycling home is not recommended as part of a cohesive strategy by the lights of C_i and u_i, so CEU-maximisation does not permit it. Of course, cycling home might maximise U_C relative to your credence and utility functions *after* having a drink. But CEU-maximisation says that what matters is carrying out optimal strategies by your initial, not current, lights.

the challenge at hand—CDT says at some time that the only U_C-maximising thing is to intend to Push, and yet Bales suggests that you may intend to do something else. So again, we have the problem that CDT's advice is not *action-guiding*: you may deviate from what CDT recommends (in virtue of the fact that you believe you will follow CDT's advice).

CEU-maximisation provides action-guiding advice in Flexibus. Say that at the outset there are three strategies: S1 (go to Damascus), S2 (first hop on Flexibus, then go to Damascus), S3 (first hop on Flexibus, then go to Aleppo). Initially, you assign high credence to Death being in Aleppo, so $C_i(\mathrm{S3} \rightarrow \mathrm{Survive})$ is greater than both $C_i(\mathrm{S1} \rightarrow \mathrm{Survive})$ and $C_i(\mathrm{S2} \rightarrow \mathrm{Survive})$. So, $CEU(S3)$ is maximal. This means that when you eventually consider changing routes and going to Damascus, CEU-maximisation will recommend persisting to Aleppo (even though at the time your reconsider, you think that persisting to Aleppo means you likely meet Death). Therefore, when Causal CEU recommends S3, you do not expect to deviate from that optimal cohesive strategy in virtue of following Causal CEU's advice. That theory therefore allows you to implement its advice.

Perhaps then if we want an action-guiding theory, we must pick between Causal CEU-maximisation and EDT. I think that EDT has some advantages over Causal CEU-maximisation.

Firstly, note that Causal CEU-maximisation will often recommend one-boxing.[158] Meacham glosses CEU-maximisation as recommending that an agent 'choose the acts she would have wanted to bind herself to'. Now, in many instances the causalist will wish that their past self had bound themselves to one-box. After all, the predictor is uncannily accurate, and had you bound yourself to one-box before their decision, they would (in all probability) have placed a million in the box. Plenty of causalists accept that if you might face a predictor in the future, you should (if you can) bind yourself to one-boxing in order to causally promote the good. Let B1 denote a plan made several days ago to one-box and B2 denote a plan made several days ago to two-box. If C_i and u_i represent your initial credence and utility functions, then if the predictor made their decision yesterday, $C_i(B1 \rightarrow \mathrm{Millionaire})$ is high and $C_i(B2 \rightarrow \mathrm{Millionaire})$ is low. So the Causal CEU of B1 is greater than that of B2. And since one-boxing is a part of B1, CEU-maximisation recommends that you one-box. Though your current choice does not influence the predictor's decision, what matters is that by the lights of C_i, *had* you bound yourself to one-boxing, that would have influenced the predictor's decision.[159]

A natural response to Causal CEU-maximisation is to wonder precisely where the motivation now is for *causalism*. Causal CEU-maximisation will agree with EDT in paradigmatic cases where

[158] I recently learned that Easwaran (Forthcoming, p. 2) makes this point.

[159] Though note that if the predictor made their decision *before* you were born, C_i will not take your binding yourself to one-box as influencing the predictor's decision. So, Causal CEU-maximisation says that you should one-box if the predictor made their decision a minute *after* you were born, but two-box if the predictor made their decision a minute *before* you were born. I think this arbitrariness speaks against CEU-maximisation—why should the precise timing of the predictor's decision make a difference to what you should do now?

CDT and EDT diverge. Perhaps we should just adopt the simpler EDT (which does not require you to keep track of complicated objects like initial credence and utility functions).[160]

Moreover, there are cases in which EDT has a clear advantage over Causal CEU-Maximisation. EDT takes your *current* conditional credences to be action-guiding, while CEU-maximisation outsources your decision to some *past* or *other self's* credences. This difference means that EDT and Causal CEU-maximisation give different verdicts in the following case (from Gibbard and Harper 1978):

> *Transparent Newcomb*: In front of you are two boxes, both transparent. One box contains a million dollars, the other a thousand dollars. The contents of the boxes were determined yesterday by an uncannily accurate predictor who placed a million dollars in the first transparent box if and only if they predicted that you take just the million (i.e., leave the thousand).

EDT says that you should two-box here. Having conditionalised on the contents of the boxes, your decision provides no further evidence and so two-boxing maximises U_E. What about CEU-maximisation? Well, consider two comprehensive strategies you could have chosen: T1 (bind yourself to one-boxing) and T2 (bind yourself to two-boxing). Which of these does best by the lights of C_i (i.e., which of them would you have wanted to bind yourself to, if you could)? Answer: T1 (binding yourself to one-boxing). After all, if before the predictor makes their decision you can tie yourself to a mast and bind yourself to one-box, then you likely cause the predictor to place a million in the opaque box. And recall the motivation behind CEU-maximisation: do the thing that you would have wanted to bind yourself to. So $C_i(\mathsf{T1} \rightarrow \mathsf{Millionaire})$ is high, while $C_i(\mathsf{T2} \rightarrow \mathsf{Millionaire})$ is low, and T1 has higher Causal CEU than does T2. Since your initial credence function has not conditionalised on what you can see, Causal CEU-maximisation recommends one-boxing *even* in some cases where you know what is in both boxes.

Note that CEU-maximisation does not require that you actually *did* at some stage plan or resolve to one-box in Transparent Newcomb. Rather, CEU-maximisation tells you that you should now do the thing that you would have bound yourself to, had you had the chance.

[160] Meacham thinks that the evidentialist should also adopt an evidentialist variant of CEU-maximisation. This is largely due to (i) a variant on the 'Why Ain'cha Rich?' argument from Arntzenius (2008), and (ii) his desire for a self-recommending decision theory. I am unsure whether a theory should be self-recommending, and see Ahmed & Price (2012) for a response to Arntzenius' WAR argument against EDT. So, I consider EDT in its standard, not CEU-maximising, formulation.

This all drives home that CEU-maximisation pays too little information to what you know about the world. Instrumental rationality for both the CDT'er and EDT'er is about choosing means that are appropriate to your ends. CEU-maximisation, however, tells you to obey the decrees of some other agent, which can mean doing something you know fails to promote your ends. Now, Meacham might take this to be a feature and not a bug of his view. He is after a theory that unifies various time-slices of agents, so he might be comfortable with one-boxing in Transparent Newcomb (just as someone like Gauthier might say that it is rational to carry out a costly threat that yields no benefit to you now, provided it was instrumentally valuable to make that threat). I personally find that verdict implausible. The causal view starts to look far less causal when we are more committed to one-boxing than the EDT'er is.

Perhaps the best way of thinking about Transparent Newcomb is as highlighting the difference between two competing accounts of rationality—the 'act-first' view and the 'strategy-first' view. The former asks which act does best by your current lights; the latter asks which kind of strategy you would like to adopt, then assess acts as parts of those strategies. Causal CEU-maximisation can deliver action-guiding advice, but only if we adopt the strategy-first view. And that requires us to shift the focus of decision theory away from what we initially took it to be—a theory about which *options* or *acts* you are permitted to choose. For those like myself who find the strategy-first view implausible (because of what it says in cases like Transparent Newcomb), only EDT provides action-guiding advice.

4.1.1.3 Time-Slices and Options

Another causalist response goes the other way. Instead of focussing on plans and commitments, we might refine our analysis of *options*. Weirich (1983) and Hedden (2012) defend a view of options as decisions, which are basic mental acts under your direct control. On Hedden's view, options must meet three criteria (2012, p. 346):

1. If something is an option for you, you must be able to do it.
2. If something is an option for you, you must believe that you are able to do it.
3. Your options supervene on your mental states (beliefs and desires in particular).

The first constraint captures an 'ought implies can' principle. The second captures what I have been assuming, that it is not action-guiding to advise you to do something you cannot do. The third is supposed to capture another important intuition behind subjective oughts and permissions: if your twin has identical beliefs and desires to you, then what is permissible for them is permissible for you. This pushes away from externalism about options and towards internalism—what counts as an option for you depends not on how the world is structured, but on the way you take the world to be.

This account suggests a simple solution to CDT's action-guiding worries. On Hedden's view, 'Going to Aleppo on the Flexibus' is not an option, since it is not something you believe that you can do (and, moreover, whether you can get to Aleppo in Flexibus depends on more than your mental states). Options, rather, are the things that you *believe* you can do (such as making a decision to go to Aleppo), so there can be no question of CDT recommending something that you do not believe you will be able to carry out.

The Weirich-Hedden view of options has some attractive features, but I think it leaves EDT with the action-guiding advantage. Let's follow Hedden (2012, p. 352) and say that options are propositions of the form 'decide to ϕ' or 'intend to ϕ'. Plausibly then, what CDT recommends initially in Flexibus is 'decide to go to Aleppo'. But now the problem for Bales' account of intentions re-arises. What does it mean to decide (or intend) to go to Aleppo *when you do not believe that you will go to Aleppo*? CDT recommends a decision to do something, but I am confident that by following CDT I will not be able to implement that decision. So again, there is something that we ordinarily say you can do (decide to get to Aleppo), that CDT judges optimal, but that you cannot do in virtue of following CDT. And again, a theory that undermines your ability to do the thing it recommends is not offering action-guiding advice.

Richter (1984) imagines a case of a CDT'er in the desert, trying to get to Aleppo or Damascus. She begins by heading to Aleppo, changes her mind and starts towards Damascus, before changing her mind again, and so on. We might expect her to reach various stages of equilibrium credence (see Joyce 2012), but as soon as she inclines towards one city over another she is no longer in equilibrium, so she reasons her way back to indecision. Here then is my question to the causalist: what is CDT recommending to this agent? It cannot be acts of the form 'Go to Aleppo' or 'Go to Damascus'. And it cannot be acts of the form 'Try to go to Aleppo' or 'Try to go to Damascus'. All CDT can say is 'just dither for a while and see what happens'—this is not far from starting, spluttering, and tying oneself up in a knot. CDT is not helping this agent to navigate the world.

Of course, the causalist might again respond that that is just the way the world is. When playing against Death, you should expect to change your mind about the choiceworthiness of what you decide, so you should expect options to be strange propositions of the form 'try to do something or other'. The causalist therefore denies that 'Decide to go to Aleppo' ever was an option for you in FlexiBus—that is just not something the world lets you decide (or intend, or try) to do. But again, that is not just the way the world is. What *makes* it the case that you cannot decide to go to Aleppo is that you follow CDT's advice. There are no forcefields or the like stopping you from carrying out what you intend. Ordinarily, we say that agents have the ability to board buses, travel to cities, and so on. And there *is* a standard of rationality that lets agents decide to do such things in cases of instability—the evidentialist standard. So, if we adopt Hedden's view of options, the outputs of decision theory for EDT resemble the kinds of things we ordinarily deliberate about, even in cases like Death in Damascus and FlexiBus. On CDT, however, the outputs of decision theory bear little resemblance to the kind of things we ordinarily deliberate about. And that is an action-guiding advantage of EDT.

So, a virtue of an action-guiding theory is that its outputs correspond to the objects of ordinary deliberation. Such a theory might finesse our ordinary talk—while 'Go to Aleppo' is not an option in the technical sense, something close enough is, 'Decide to Go to Aleppo'. But CDT goes one step further and undermines our ability to even *decide* or *intend* to go to Aleppo in FlexiBus. Since *I* want a theory to help me get around the world, I prefer EDT.

4.2 <u>Don't Get Framed: CDT and Framing Options</u>

Claim: a theory is not action-guiding if its advice varies with equally legitimate descriptions of a decision. And CDT's advice varies with equally legitimate descriptions of a decision.

To see the problem for CDT, consider the following case from Ahmed (2014a):

> *Dicing with Death*: You are deciding whether to go to Aleppo or Damascus. As always, Death is a highly accurate predictor of your decisions (Death gets things right 90% of the time), and you know that your decision is causally independent of Death's. This time you have the added option of paying a dollar to toss a coin that teleports you to Aleppo on landing Heads and Damascus on landing Tails—Death is no better than chance at predicting how this coin lands.

We can represent this:[161]

[161] Note again the assumption (which might be a convenient fictionalisation) that there is a counterfact of the matter as to how the coin would land if tossed.

	Death in Damascus & Heads	Death in Damascus & Tails	Death in Aleppo & Heads	Death in Aleppo & Tails
Go to Aleppo	10	10	0	0
Go to Damascus	0	0	10	10
Dice	9	-1	-1	9

Provided $C(\text{Heads}) = C(\text{Tails}) = .5$, we get that $U_C(Dice) = 4$. If $C(\text{Death in Aleppo}) = C(\text{Death in Damascus}) = .5$, then $U_C(Aleppo) = U_C(Damascus) = 5$, so CDT recommends going to either Aleppo or Damascus over Dicing. More generally, whatever your credences either $C(\text{Death in Aleppo}) \geq .5$ or $C(\text{Death in Damascus}) \geq .5$, so at least one of $U_C(\text{Aleppo})$ or $U_C(\text{Damascus})$ is greater than 5. So, CDT *never* permits Dicing in Ahmed's case.

Ahmed claims that you should Dice, so he takes this case to be a counterexample to CDT.[162] I do not think that it is. Say that the causalist is confident that Death is in Aleppo—then they think they likely avoid Death by going to Damascus, and Dicing would increase their chances of meeting Death. So it is a mistake to Dice, simply because it is mistake to take a reasonable chance of death when you can likely avoid Death. Moreover, CDT says *initially* that Dice is worse than both Aleppo and Damascus. But the CDT'er accepts that on choosing a city, *that* decision looks worse than Dicing. So, we can employ the same debunking argument here as in standard cases of instability—if it seems like CDT is offering the wrong advice, that may just be because we judge CDT's pre-choice advice by our credences on deciding to implement that advice.

My interest, however, is not with Dicing with Death as a counterexample to CDT. My interest is in the lesson Ahmed draws from the case. He claims that:

> 'Certainly [dicing] looks like a good offer. Would you rather be playing hide-and-seek against (a) an uncannily good predictor of your movements or (b) someone who can only randomly guess at them?' (Ahmed 2014a, p. 589)

He later concludes:

> 'Platitude: in a game that you lose iff your causally isolated opponent correctly predicts your act, you are better off playing against a hopeless predictor than against a very good predictor.' (Ahmed 2014a, p. 592)

[162] A similar case is discussed as a counterexample to CDT by Spencer and Wells (2019).

I agree—it is close to nonsensical to say that you are better off playing against an accurate, malevolent predictor than an inaccurate, malevolent one. Claim: CDT vindicates this platitude.

Ahmed's platitude is not about what you should choose in Dicing with Death, but which *kind* of predictor you should want to play against. Say that you are choosing to play hide-and-seek with either Death (who guesses right 90% of the time) or Death's hopeless brother Seth (who is no more accurate than chance). Now, before choosing a predictor you have high credence in the claim that 'If I were to play against Death, I would die'—the chance of your meeting Death if you play against Death is 90%. Similarly, you assign $.5$ credence to the claim that 'If I were to play against Seth I would die'—the chance of your meeting Seth if you play against Seth is $.5$.[163]

So, the relevant credences in act-outcome non-backtracking counterfactuals are:

$$C(\text{Play Death} \rightarrow \text{Die}) = .9 \text{ and } C(\text{Play Seth} \rightarrow \text{Die}) = .5$$

We now calculate:

$$U_C(\text{Play Death}) = C(\text{Play Death} \rightarrow \text{Die}) \cdot 0 + C(\text{Play Death} \rightarrow \text{Survive}) \cdot 10$$

$$= 1$$

$$U_C(\text{Play Seth}) = C(\text{Play Seth} \rightarrow \text{Die}) \cdot 0 + C(\text{Play Seth} \rightarrow \text{Survive}) \cdot 10$$

$$= 5$$

So, CDT recommends playing against the poor predictor over the good one, vindicating Ahmed's platitude. There is nothing mysterious in this: you would do better by selecting the hopeless predictor over the competent one, so playing against Seth causally promotes the good.

Let's return to Dicing with Death. What if you frame your decision not as a decision about where to go, but a decision about how to decide? That is, you might wonder whether you should select a city by *deliberating* or *randomising* (by tossing a coin). The relevant credences in causal, non-backtracking counterfactuals are now:

$$C(\text{Deliberate} \rightarrow \text{Die}) = .9$$

$$C(\text{Dice} \rightarrow \text{Die}) = .5$$

[163] This is not to say that your choice of where to go influences Death prediction—they are causally isolated from you. Rather, you know that *you* are disposed to go where Death is, so when you take an external perspective on your decision, you think it likely that you will go where Death is.

When you think about your decision in this way, you think of yourself as doomed—you treat your possible deliberation as a chance process that is biased towards death. On the basis of the above counterfactuals, we now calculate $U_C(\text{Deliberate}) = 1$, while $U_C(\text{Dice}) = 5$. So, CDT says that you should Dice rather than Deliberate—the intuition that Ahmed endorses.

We now have a puzzle re-emerging. Framed one way (as a choice between Aleppo, Damascus, and Dicing), CDT tells you *not to randomise*. Framed another way (as a choice between deliberating or randomising), CDT tells you *to randomise*. So, does CDT tell you to Dice or not?

There seem to be two legitimate ways of framing the decision you find yourself in. One framing we might call the 'first-person' stance, in which you think about each act you might perform and its instrumental utility. The other framing we might call the 'third-person' stance, in which you think about the instrumental utility of *your deciding*, as opposed to say tossing a coin. Both of these strike me as legitimate framings of the decision you find yourself in—you can think about specific acts (like choosing a certain city) *and* you can think about choice methods (like deliberating versus randomising). Yet because of the way that CDT calculates instrumental utility, it yields different verdicts across these two framings.

In general, I claim that EDT does not exhibit this kind of insensitivity to framing. We can see this by considering EDT's advice in Dicing with Death, then drawing a general lesson about how EDT values deliberation over an option set. Because EDT evaluates each option based on the news it provides, EDT's evaluation of options in a set is consistent with its evaluation of *deliberating over that set*.

In Dicing with Death, consider how EDT evaluates Going to Aleppo and Going to Damascus:

$$U_E(\text{Aleppo}) = C(\text{Die}|\text{Aleppo}) \cdot 0 + C(\text{Survive}|\text{Aleppo}) \cdot 10$$

$$= .9 \cdot 0 + .1 \cdot 10 = 1$$

$$U_E(\text{Damascus}) = C(\text{Die}|\text{Damascus}) \cdot 0 + C(\text{Survive}|\text{Damascus}) \cdot 10$$

$$= .9 \cdot 0 + .1 \cdot 10 = 1$$

Crucially, your conditional credences reflect your views about *how likely the predictor is to be correct*. That is $C(\text{Die}|\text{Aleppo}) = C(\text{Die}|\text{Damascus}) = C(\text{Predictor Correct})$. So, another way of thinking about the decision is that we calculate utility relative to the probabilistically act-independent states 'Predictor Correct' and 'Predictor Incorrect':

$$U_E(\text{Aleppo}) = U_E(\text{Damascus}) =$$

$$C(\text{Predictor Correct}) \cdot 0 + \text{C}(\text{Predictor Incorrect}) \cdot 10$$

U_E evaluates each option based on how likely the predictor is to be correct. How then does EDT evaluate your *deliberating over the set* $\{\text{Aleppo, Damascus}\}$? Again, we think about the probability that Death has guessed correctly if you deliberate over this set, so we simply calculate:

$$U_E(\text{Deliberate}) = C(\text{Predictor Correct}|\text{Deliberate}) \cdot 0 + \text{C}(\text{Predictor Incorrect}|\text{Deliberate}) \cdot 10$$

Again assuming that Death's accuracy is probabilistically independent of what you now decide, we get:

$$U_E(\text{Deliberate}) = C(\text{Predictor Correct}) \cdot 0 + \text{C}(\text{Predictor Incorrect}) \cdot 10$$

And this is just the same quantity as $U_E(\text{Aleppo})$ and $U_E(\text{Damascus})$. When evaluating options like 'Go to Aleppo' and 'Go to Damascus', EDT tells you to think about the probability that the predictor is correct. And when you deliberate over those options, EDT also tells you to think about the likelihood that the predictor is correct. So, EDT delivers consistent verdicts in Dicing with Death, regardless of whether you think first-personally (i.e., about the utility of each option in $A = \{\text{Aleppo, Damascus}\}$) or third-personally (i.e., about the utility of choosing from $A = \{\text{Aleppo, Damascus}\}$).

To see EDT's insensitivity to framing more generally, we need to be able to talk about the utility of deliberating over an arbitrary option set $A = \{a_1, \dots, a_n\}$. I assume that rational agents believe they will deliberate rationally, which for the evidentialist means they believe that they will select a U_E-maximising option in A as the result of deliberation. Let such a U_E-maximising option be m. Above, we worked with two probabilistically act-independent states that determined the outcome of your choice, 'Predictor Correct' and 'Predictor Incorrect'. But in general, we can take a set of dependency hypotheses $\{P_1, \dots, P_n\}$ such that each P_i determines the outcome of your choice and is probabilistically independent of anything you might do. Heuristically, we can think of each P_i as some possible predictor—in Dicing with Death we let $P_1 = \text{Correct Predictor}$ and $P_2 = \text{Incorrect Predictor}$.

When the evidentialist thinks third-personally about the utility of deliberating over A, denoted $Del(A)$, they simply reason as they did in Dicing with Death: they consider the various 'predictors' they might be facing, and think about what the consequence of deliberating is if each P_i holds. So, letting $m \in A$ be a U_E-maximising option in A, we calculate:

$$U_E\big(Del(A)\big) = \sum_i C(P_i|Del(A)) \cdot u(o_{m,P_i})$$

And again, since which P_i holds is probabilistically independent of what you decide to do, we get:

$$U_E\big(Del(A)\big) = \sum_i C(P_i) \cdot u\big(o_{m,P_i}\big)$$

$$= \sum_i C(P_i|m) \cdot u\big(o_{m,P_i}\big)$$

$$= U_E(m)$$

Essentially: the U_E of *deliberating over* A just is the U_E of *a maximal option in* A. So, EDT ranks 'deliberating over A' above some act just in case the maximal option in A is preferable to that act. In particular, if we let r_A be some chance distribution over the options in A, then $U_E\big(Del(A)\big) \geq U_E(r_A)$ just in case you prefer some option in A to r_A.

Contrast this with the causalist, who might prefer each option in A to r_A, though prefer r_A to deliberating over A—witness Dicing with Death. From the third-person perspective, the causalist thinks not about patterns of causal dependence between *acts* and *outcomes*, but the expected causal effects of *deliberating itself.* This means that they evaluate, say, the *option* 'Go to Aleppo' relative to the dependency hypotheses $K_1 = \text{Death in Aleppo}$ and $K_2 = \text{Death in Damascus}$. But when the causalist shifts perspective and asks what the consequences of their own deliberation would be, the predictor's success rate matters, and they calculate the instrumental utility of deliberating relative to $P_1 = \text{Correct Predictor}$ and $P_2 = \text{Incorrect Predictor}$. And the lesson from Dicing with Death is that these perspectives yield inconsistent verdicts.

Now, this is not quite a contradiction for the causalist. Formally, a decision situation specifies an option set, so there is no *single decision situation* in which CDT tells you, say, to Dice and not to Dice. What is true is that when there are multiple ways of describing your options, CDT's advice is sensitive to how you describe your options.

Recall that I want a decision theory to help me navigate the world. Yet CDT gives me inconsistent advice based on how I frame my options. Whatever I do therefore, I know that there is some perspective from which CDT tells me that I have done the wrong thing. And the fact that CDT recommends *this* rather than *that* is due to an arbitrary choice on my part to describe my options in a certain way. So, I cannot look to CDT to make a single, unambiguous recommendation. I therefore must settle with overriding CDT's advice relative to some legitimate perspective I might take on the decision I face.

Objection: There is a natural response to this kind of problem—insist on a single, privileged framing.

Response: Our decision theory should be responsive to the way that agents themselves view the decisions they face. We *do* in fact deliberate about how to decide, so it would be a serious shortcoming if our theory could not place a value on deliberation itself. EDT, but not CDT, allows us to value decisions in a way that is consistent with the value of the options that make up those decisions.

Note also that insofar as people have insisted on a single, privileged framing of options, it is usually taken to be the *narrowest* framing—the most fine-grained partition of the things you can bring about such that you cannot 'do anything that implies but is not implied by a proposition in [that] partition' (Lewis 1981a, p. 7). This might rule out acts like 'Deliberate over A', since both 'Go to Aleppo' and 'Go to Damascus' imply that you deliberate over $\{\text{Aleppo}, \text{Damascus}\}$, but not vice versa. 'Deliberate over A' is too coarse-grained to be in an option set.

But the question is why we *should* insist on the narrowest framing. Lewis stipulates that it is the right framing, but it is somewhat unclear why he takes more coarse-grained framings to be illegitimate descriptions of a decision. Sobel (1988d, p. pp. 203-204) does provide an argument. Firstly, he considers acts of the form 'do f or g'. He claims that we can rarely realise these 'disjunctive acts' as options—we are in a position to enact f, we are in a position to enact g, but we are not in a position to enact the disjunctive 'f or g'. That might be right, but we surely *can* deliberate about whether to perform one of f or g. We do know what it is to deliberate between options (we do it all the time), and an agent facing a decision might ask themselves what the utility of deliberating between f and g is. So, even if Sobel is right that 'disjunctive options' are mysterious, this does not speak in favour of eliminating acts like 'Deliberate between f and g' from our option set.

Sobel then claims that we should simply set the value of 'do f or g' (or, we might say 'deliberate between f and g') to the value of the preferable disjunct, since the disjunctive act 'f or g' is 'made up' of the narrower acts f and g. So, we do not directly assign instrumental utilities to coarse-grained acts like 'deliberate about which city to go to'; rather we simply look at the fundamental utilities of narrower options (acts like 'Go to Aleppo' and 'Go to Damascus'). But this does not seem to be an *argument* that we should adopt the narrowest partition of options. Sobel is claiming that the value of a coarse-grained option (say, a decision to decide between more fine-grained options) should be identified with some utility of a narrower option that makes that coarse-grained option up. But this is not an *argument* for a particular framing so much as a statement of the intuition that the value of deliberation ought not exceed the value of the options that you deliberate over. And that is precisely the property that EDT has but that CDT lacks. Sobel is really saying that it would be odd to rank, say, 'go to some city' above going to any particular city, since the former is made up of the latter. But that says nothing about the former not being a legitimate object of deliberation—it is a judgement that the value of the former *ought* not exceed the value of the latter. And we have already seen in Dicing with Death that it is EDT, not CDT, that vindicates this judgement.

So, I see no reason to insist that option sets be the narrowest available partition of the things you might do. We can reasonably ask what the value of deliberating over A is, and an action-guiding theory should offer advice that is insensitive to how you choose to frame your options.

Another prominent account of optionhood, the Weirich-Hedden view, which I have already discussed, will not help the causalist either. Say that options are basic mental acts—decisions (or intendings, tryings, etc.). While this identifies options with decisions, it does not specify what the contents of those decisions are. An agent could take their options to be 'Decide to go to Aleppo, Decide to go to Damascus, or Decide to Dice', but they could just as well take their options to be 'Decide to deliberate, or Decide to randomise over destinations'. And again, CDT will offer different advice depending on which framing we choose, while EDT delivers consistent advice across framings.

Objection: You might think that when I described Dicing with Death in terms of choice methods (deliberate or dice), I changed the case. CDT simply recommends an optimal option in each decision situation—we just have to be careful not to change the decision situation.

Response: It is true that CDT recommends a single option in any decision situation. So, from a formal perspective, you might simply say that the option sets $\{\mathsf{Aleppo}, \mathsf{Damascus}, \mathsf{Dice}\}$ and $\{\mathsf{Deliberate}, \mathsf{Dice}\}$ represent two distinct decision situations. And since Ahmed describes his case with the former, not the latter, you might think we should only concern ourselves with CDT's advice relative to the former.

But nature does not come with decisions neatly labelled as this-or-that formal object. If *you* face Dicing with Death, *you* have multiple ways of deliberating about that case. It is not as if you can look to the heavens and see 'Deliberate about which city to go to, *thou shalt not* think about the utility of deliberating-versus-randomising'. So even if from a formal perspective CDT offers unambiguous advice in each decision situation, when it comes to real agents making real choices, they might be torn about which formal 'decision situation' models their predicament.

Of course, in *some* situations there might be a single way of framing the agent's options that stands out. Perhaps you are presented with a choice of cities to travel to, and you simply find yourself deliberating with no ability to think about more coarse-grained options. And perhaps in another situation you are presented with a choice of cities to travel to, then asked explicitly whether they would like to randomise or choose yourself. But Dicing with Death does not fit either of these moulds. The CDT'er might ask themselves what the value of deliberating is (*low*, so Dice and stop yourself from deliberating!) or they might ask themselves what the value of, say, Aleppo is (*high*, so go there and stop yourself from Dicing). Both correspond to ordinary forms of practical reasoning.

Objection: CDT does offer action-guiding advice: it is *indeterminate* what you should do in Dicing with Death.

Response: If we accept that what you should do in Dicing with Death is indeterminate, this is still puzzling. It is hard to know what to do with indeterminate advice—it always tells you that you are, in a sense, doing the wrong thing. Perhaps it is not *determinately* the case that you do the wrong thing by Dicing in Dicing with Death; nonetheless, there is a legitimate perspective from which you do the wrong thing. Whenever I ask CDT whether *this* is the thing to do, it responds 'in a way yes, in a way no'. So what am I supposed to *do* then?

Now, sometimes a theory's advice might be indeterminate because of worldly indeterminacy. If you are betting on whether some cloud is larger than 10 cubic meters, or if you are choosing whether to read Shakespeare or listen to Bach, then maybe there is no fact of the matter what you should do. And it makes sense that decision theory leaves you in a bind here—whatever you do (prioritise Shakespeare's eloquence over Bach's intricacy, or vice versa), there is *some* sense in which you are doing the wrong thing. But Dicing with Death is not like that. Every outcome has a determinate utility, and you have precise credences. So indeterminate advice is not tracking some indeterminate feature of the world. Rather, because CDT is sensitive to framing, and there is no fact of the matter as to how you should frame your decisions, CDT makes multiple, inconsistent verdicts. EDT, however, *does* provide consistent verdicts across framings. So, the indeterminacy is not in the world, but in the way that CDT advises you to reason about the world. We need not be dragged by the causalist into saying 'that's just how the world is'.[164]

So, CDT leaves an important part of your decision up to you. Some parts of your decision surely are up to you—your risk-attitudes, your tastes, and so on. But of course a theory's advice should be sensitive to those features of your psychology which represent your ends and how you assess means to ends. But framing is not like that. It is not as if each agent has a 'framing attitude' that reflects how they evaluate means to ends. Each agent might wonder about different framings and take those framings to represent equally legitimate descriptions of the decision they face. A theory whose advice is sensitive to choice of framing does not, therefore, offer consistent advice in the one situation. EDT's advice is robust across framings and so provides a single, unambiguous verdict that I can implement.

4.3 <u>Conclusion: Standoffs and Picking a Side</u>

I began this chapter by asking what we want a decision theory *for*. I then noted that even if 'rationality' underdetermines whether we should adopt CDT or EDT, we might still judge one of those theories to be more fit for purpose. (Just as we might judge some standards of baldness to be more fit for purpose.)

I want a decision theory to provide action-guiding advice—the kind of advice that I can use to navigate the world. Such a theory will recommend things that I can implement by following that theory, it will recommend things that correspond to the ordinary objects of deliberation, and it will offer robust advice across legitimate framings of the same decision. In the cases I have discussed, EDT has these properties, but CDT does not. So, as we follow Horgan in constructing a theory that vindicates what already seems right to us, I find that EDT vindicates more plausible features of our ordinary practical reasoning than does CDT. Implicit here is that we are not judging theories based just on which 'brings about the best results'—rather, we judge

[164] Williams (2014) and Bales (2018c) defend views on which if both f and g are indeterminately permissible, then both are choiceworthy. If so, then perhaps the causalist might say that both Dicing and Not Dicing are permissible in Dicing with Death (since both are indeterminately permissible). I do not go this route—I am unsure what it means to say that something is determinately choiceworthy but indeterminately permissible. Note that in the Introduction I simply defined a choice function as mapping an option set to a subset of *permissible* options. This is because I doubt there are two concepts here—permissibility and choiceworthiness—in the first place: if you tell me that something is determinately choiceworthy, then that just means that there is a fact of the matter about whether I am permitted to choose it (I am). So, though it deserves a fuller discussion, I side with Rinard (2015), who thinks that permissibility and choiceworthiness go together (in particular, indeterminate permissibility implies indeterminate choiceworthiness).

theories based on the desirable features that they have. Again, if we are looking for 'the' correct theory of rationality, then maybe the causalist and evidentialist are locked in a stalemate. But if we are each free to construct a theory that seems right to us, I see no reason *not* to take as an inputs the features we want from a theory of instrumental rationality, like action-guidingness.

What about you? Perhaps you want something different from a decision theory. Or perhaps you find yourself unable to shake the two-boxing verdict in Newcomb's Paradox, so you simply accept the quirks of CDT that I have discussed. I do not think that you are wrongheaded—CDT is a perfectly coherent standard of rationality, and we can chalk up CDT's action-guiding quirks as features, not bugs, of rationality when playing hide-and-seek against Death (from *your* perspective, framing indeterminacy and a failure of action-guidingness just reflect the way the world is).

But I am not playing the internal consistency game. I find myself genuinely torn in Newcomb's Paradox, and I can get myself in both the causalist and the evidentialist frame of mind. But EDT rationalises more of what I want from a theory of instrumental rationality—consistent advice across times and presentations of the same choice. I take these to be theoretical virtues of an action-guiding theory—I can understand and implement EDT's advice in a way that I cannot understand and implement CDT's advice. And that is how I became a one-boxer. (Take note, any predictors who might be standing by with a spare million.)

Chapter 5

Dominance and Intransitivity

5.0 Cyclic Preferences

So far, I have assumed that preferences over acts satisfy *Transitivity*: if $a \succ b$ and $b \succ c$, then $a \succ c$ (similarly if $a \succcurlyeq b$ and $b \succcurlyeq c$, then $a \succcurlyeq c$).[165] This is a widely accepted structural assumption. Indeed, since the greater-than relation is transitive on the real numbers, the traditional decision-theoretic project of assigning each outcome a real-valued utility presupposes Transitivity.

Nonetheless, I am sceptical of Transitivity. Unlike many other decision-theoretic principles, Transitivity constrains not just which means are appropriate to take to your ends, but your ends themselves. Independence, Betweenness, State-wise Dominance, and so on each take your preferences over ends (i.e., outcomes or degenerate acts) as given. They then say which preferences over means (i.e., non-degenerate acts) are appropriate given those more basic facts about ends. This fits with a broadly Humean picture of decision theory—we are not in the business of dictating tastes, desires, or goals. Contrast this with Transitivity, which does place substantive constraints on your ends. While it is permissible to prefer Bach over Bob Dylan, to prefer Bob Dylan over Beethoven, and to prefer Beethoven over Bach, Transitivity forbids holding each of these preferences simultaneously. This places a substantive constraint on your ends themselves and so says that certain patterns of tastes, desires, or goals are impermissible. The Humean in me resists this.[166]

In this chapter and the next therefore, I investigate dominance principles in the absence of Transitivity. I focus especially on the case of cyclic preferences over outcomes—preferences of the form $x > y > z > x$. In Section 5.1 I discuss Money Pumps, both diachronic and synchronic, and argue that they fail to undermine rational cyclic preferences. In Section 5.2 I present a challenge involving risk, the *Easy Lottery*, which shows that cyclic preferences are

[165] Since Transitivity constraints preferences over degenerate acts, this imposes Transitivity on preferences over outcomes as well.

[166] Some views of preference may render intransitive preferences nonsensical (e.g., if as in Bradley 2017, Section 4.2 preferences are a certain kind of judgement of betterness and the 'better than' relation is necessarily transitive). On such views, even the Humean should accept Transitivity. Here I assume that whatever preferences are, we can make sense of cyclic preferences. For example, cyclic preferences are intelligible if preferences are judgements of betterness and the 'better than' relation admits cycles, or if preferences supervene on facts about what you have most subjective reason to choose (cf. Thoma 2021, Section 7) and the 'more reason than' relation admits cycles (as is permitted by Snedegar 2017).

incompatible with at least one plausible dominance principle. I do not take a stand on which dominance principle to reject (or whether we should reject cyclic preferences), but I show how each response to the Easy Lottery is vindicated by some formal rule for comparing acts whose outcomes are not transitively ordered. In Chapter 6 I turn to the general choice problem, which is how cyclic preferences guide choice in option sets containing more than two options; I canvas various proposed solutions and suggest my own rule, *Least Bad*, which delivers plausible verdicts in a range of cases. Least Bad, however, violates a minimal dominance constraint on choice, which is that rational choices are not strictly dispreferred to every available alternative. I suggest that this is a feature, not a bug.

5.1 Intransitivity: A Brief Survey[167]

If you thought only about straightforward cases, you might take Transitivity to be well-motivated. After all, you prefer $\$100$ to $\$10$ and $\$10$ to $\$1$, so you prefer $\$100$ to $\$1$. Similarly, you might prefer Bach to Bob Dylan and Bob Dylan to Cher, then straightaway conclude that you prefer Bach to Cher. But these are easy cases. If more money is preferred to less, then we need not appeal to *Transitivity* to explain your preferring $\$100$ to $\$1$—we can simply note that $\$100$ is a greater sum of money than $\$1$. And to explain that Bach is preferred to Cher, we can simply cite the fact that Bach is one of the great composers and, well, Cher is not.

If you were to start thinking about less straightforward cases, then Transitivity might appear less well-motivated. I am sure that I prefer Bach to Bob Dylan and Bob Dylan to Beethoven, but it is unclear that I prefer Bach to Beethoven. Indeed, I find myself choosing to listen to Beethoven far more than Bach when browsing classical music—I really do seem to prefer Beethoven to Bach. And I do not think these preferences are mistaken.[168]

[167] A few terminological points. (i) Sometimes Transitivity is formulated as: if $a \succ b$ and $b \succ c$, then $\neg(c \succcurlyeq a)$. This formulation is equivalent to the one in the main text given completeness (i.e., the assumption that you have a preference between any two acts). Failures of completeness give rise to their own interesting puzzles, so to bracket them and highlight the distinctive puzzles generated by failures of Transitivity, I assume completeness. (ii) Non-transitivity refers to any failure of transitivity, while I take *cyclic* preferences to be preferences of the form $a \succ b \succ c \succ a$. So, preferences of the form $a \succ b \succ c \sim a$ are non-transitive, but not cyclic.

[168] A separate question is how widespread failures of transitivity are. Tversky (1969) is the *locus classicus* for demonstrating the empirical failure of transitivity, though Regenwetter et al. (2010, 2011) challenge Tversky's interpretation of the results and much of the literature following Tversky. While evidence of widespread cyclicity might support the claim that such preferences are rational, we need not show that cyclicity is widespread to show that it is rational. (I do note, however, that the methodological points raised by Regenwetter et al. may put pressure on the intuitions about cases often cited to support rational cyclic preferences. Assessing their methodology and its implications is for future work.)

Thinking about still harder cases puts more pressure on Transitivity. For example, say that you care about three kinds of thing when choosing between options: health, wealth, and wisdom. You are considering three career choices—Sports Star, CEO, and Philosopher—that score differently on each dimension:

	Health Rank	Wealth Rank	Wisdom Rank
Sports Star	1	2	3
CEO	3	1	2
Philosopher	2	3	1

These rankings are self-explanatory: being a Sports Star requires a lot of health, generates a good amount of wealth, but leaves you with little time for intellectual pursuits; being a CEO generates an enormous amount of wealth, requires some wisdom, but the wining-and-dining severely impacts your health; Philosophers become extremely wise, get a moderate amount of exercise, but they end up poor.[169] It seems rational to prefer a career as a Sports Star to a career as a Philosopher (Sports Star does better on two out of three dimensions), a career as a Philosopher to a career as a CEO (Philosopher does better on two out of three dimensions), and a career as a CEO to a career as a Sports Star (CEO does better on two out of three dimensions). This seems to be a reasonable way of forming preferences, yet it gives rise to intransitivity.[170]

Intransitive preferences also appear reasonable in so-called spectrum scenarios. One such case comes from Quinn (1990):

> *The Self-Torturer*: Jane is in no pain (call this initial pain-level p_0). For any pain-level p_i, she prefers to increase her pain by some fixed, barely-perceptible increment, ϵ, and receive $\$1{,}000$ over staying at her current pain-level. This means that Jane has preferences:
>
> $$p_n + \$n \cdot 1{,}000 > p_{n-1} + \$(n-1) \cdot 1{,}000 > \cdots > p_1 + \$1{,}000 > p_0$$

[169] Partly because of the salary, partly because they are all two-boxers.

[170] Or you might reason conversely: since the method gives rise to intransitivity, it is not a reasonable way of forming preferences. This touches on another methodological question—how to weigh off judgements about cases against judgements about principles. I here adopt a *permissive methodology*, which counts apparently reasonable preferences as rational and only rejects the rationality of those preferences in light of a compelling argument for doing so.

But p_n is an excruciating level of pain, and Jane prefers poverty with no pain to riches with excruciating pain. So, while Jane always accepts a slight increase in pain for substantial increase in gain, she still (seemingly rationally) prefers no pain and no gain to excessive pain and excessive gain. Jane has cyclic preferences, though again each of her preferences seems rational.[171]

5.2 Money Pumps

This is not the place to review every argument for or against rational intransitivity—see Anand (1993, 2009) for a summary and critique of the main arguments against. I here focus on various *Money Pump* arguments against cyclic preferences. This is because such arguments appeal to various dominance principles, which is my focus here.

Before getting into the arguments themselves, recall the following terminology: for an agent with preferences over outcome set O, x is *absolutely preferred* to y, denoted $x \gg y$, if (i) $x > y$, (ii) for any $z \in O$, if $z > x$ then $z > y$, and (iii) for any $z \in O$, if $y > z$ then $x > z$. If the strict preference relation admits cycles, then it is unclear that one outcome's being strictly preferred to another means that it is an improvement all things considered. The absolute preference relation is stronger than strict preference and is acyclic,[172] so I take it to capture one sense in which an outcome can be an improvement (all things considered) over another given cyclicity.

5.2.1 The Diachronic Money Pump

[171] Or again, you might think that because Jane has intransitive preferences, she must be making some mistake. Perhaps this is a kind of Sorites case, and we should say that one of Jane's preferences is irrational, though it is hard to say which. Voorhoeve and Binmore (2006) have another proposal: we rely on *similarity-based reasoning*. That is (following Rubinstein 1988), when two outcomes differ greatly in one respect and minimally in another, we ignore the respect in which they minimally differ; this is typically innocuous, though can add up to irrationality when repeated in cases like the Self-Torturer. I am unpersuaded by Voorhoeve and Binmore's point. For one, even if we engage in similarity-based reasoning, it is far from obvious that such reasoning is irrational. When two outcomes differ minimally in one respect, why *not* form a preference by considering only the respect(s) in which they differ greatly (perhaps some differences are minimal enough to be irrelevant, even from the perspective of an ideally rational agent)? More generally, it strikes me as plausible that we are rational to treat pain non-additively: when considering a great increase in pain, we need not treat it as the sum of several small increases. We really need an argument that we *ought* to treat pain additively, while pointing to similarity-based reasoning merely provides an explanation for why we do not. Again, a permissive methodology will treat this case is defeasible evidence for rational intransitivity.

[172] Proof: Say that $a \gg b \gg c$. If $c \gg a$, then for any x such that $a > x$, $c > x$. Since $a > b$, we get $c > b$, contradicting that $b \gg c$. So $\gg$ is acyclic.

The Diachronic Money Pump shows that agents with cyclic preferences knowingly make sequences of choices that leave them worse off than they need be. But plausibly, rational agents never knowingly end up worse off than they need be, so cyclic preferences are not rational. Recall from the Introduction that one plan *absolutely dominates* another if it leads to an absolutely dispreferred outcome, regardless of how the world turns out to be. That is:

> **Non-Absolutely Dominated Plans:** A rational plan is one that does not lead to an absolutely dominated outcome in every state-by-state comparison to some alternative available plan.

The Money Pump shows that agents with cyclic preferences choose dominated plans, so by the Non-Absolutely Dominated Plans constraint they behave irrationally.

To illustrate, I use a version of the Money Pump due to Cantwell (2003).[173] Cantwell's argument, like all diachronic Money Pumps, leverages a substantive assumption about the structure of your preferences, that:

> **Availability of Small Fees:** For any outcomes o, o' such that $o > o'$, there is some outcome $o - \$\epsilon$ such that $o > o - \$\epsilon > o'$, and $o \gg o - \$\epsilon$.

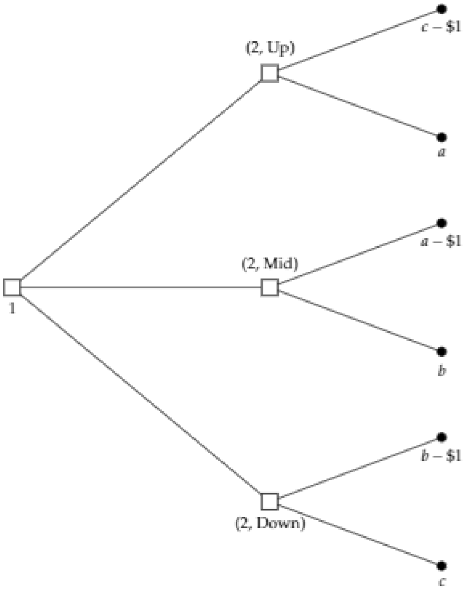
This says that for each pair of outcomes, there is some small fee that you always absolutely prefer not to lose, but that you are prepared to pay to satisfy your preferences.[174] (Note that the size of $\$\epsilon$ may vary depending on o and o'.) Given cyclic preferences $a > b > c > a$, recall from the Introduction Cantwell's (2003, p. 389) case: [175]

> *Money Pump*: You may select any one of $\{a, b, c\}$. After you make your decision, I will offer you one more choice. If you chose a, I offer you c for $\$1$. If you chose b, I offer you a for $\$1$. If you chose c, I offer you b for $\$1$. That is:

[173] I focus on Cantwell's Money Pump because it is far simpler than many in the literature. Everything I say about Cantwell's case, however, applies to other diachronic Money Pumps.

[174] Since $o - \$\epsilon > o'$, you choose $o - \$\epsilon$ in a pairwise choice between them.

[175] In what follows I assume that the small fee that you pay to satisfy your preferences is $\$1$. This is to simplify the case—we could let the amount you are prepared to pay vary between each pair of outcomes.



To see that you violate the Non-Absolutely Dominated Plans constraint, assume that you choose in line with your preferences at the second node (given that each choice you face at the second node is pairwise, you choose the preferred option from the pair). This means that at node 2 you select the 'Down' option. You therefore end up with one of $\{a-\$1, b-\$1, c-\$1\}$. But there are feasible sequences that yield each of $\{a, b, c\}$—simply select the 'Up' option at node 2. So, following your cyclic preferences means you are guaranteed to end up worse off than you need be: whichever outcome you end up with, you could have saved a dollar.[176]

[176] Ahmed (2017, Section 6.2) raises two objections to Cantwell's Money Pump: (i) So long as $a-\$1, b-\$1, c-\$1$ are all better than the status-quo, this 'Money Pump' does not leave you worse off than you started, and (ii) No cunning bookie has an incentive to offer you this Money Pump since they are made no better off by it. If Ahmed's criticisms are successful, note that what I say below applies to traditional Money Pumps that avoid Ahmed's critique. Nonetheless, I do not think that either objection undermines Cantwell's argument. In response to (i), I take the alleged problem with cyclic preferences to be that they make you worse off than you need be, not that they make you worse off than you were at some time. If we criticise an agent, it is for failing to do as well as they could have. (Analogy: Consider an agent with complete, transitive preferences choosing from a finite option set. If such an agent pays a dollar for an option they could have for free, they are irrational in virtue of throwing away money—it does not matter whether the option they pay for leaves them better or worse off than some status quo.) In response to (ii), the Non-Absolutely Dominated Plans constraint does not mention the probability that you face the decision situation in question—that there exists a possible situation where you lose a dollar is enough. Moreover, we can think of plenty of situations where a bookie would be incentivised to offer you Cantwell's Money Pump. For example, assume that your cyclic preferences $a > b > c > a$ are insensitive to changes in your background wealth. The case could then involve the bookie first *selling* you any one of $\{a, b, c\}$ for its maximum fair price, then offering to trade for a dollar—the result is that the bookie sells one of $\{a, b, c\}$ for a dollar more than is fair. Or say that the exploiter owes you one of $\{a, b, c\}$—they can then offer you Cantwell's Money Pump to make a quick buck out of you, leaving you worse off than you need be. Or we can imagine situations with multiple bookies: say that you get to choose one of $\{a, b, c\}$—if you choose a (b, c respectively), then anyone standing by with c (a, b respectively) can make a quick buck out of you by offering a trade.

There have been a range of responses to the Money Pump. Many focus on whether you really are exploitable in the money pump (I discuss this further in Chapter 6), others on whether the sequences that save you a dollar are really *feasible* alternatives.[177] I adopt a different approach here and argue that even if cyclic preferences make you worse off than you need be in the Money Pump, this does not show that those preferences are irrational.

I begin by asking precisely why it matters that you choose a dominated sequence of acts. The primary goal of decision theory is to recommend optimal *options*, and it would be a substantive assumption that every sequence of optimal options is itself an optimal sequence. Indeed, if you are motivated to accept cyclic preferences, this is just the assumption you might reject. To get at this idea, consider that cyclic preferences can be rationalised by *essentially comparative* or *context-sensitive evaluations*—your attitudes, reasons, or evaluations depend crucially on what each outcome is compared to. For example, in Quinn's Self-Torturer case, the complex way that pain- and wealth-increases interact means that there is no context-free way of assessing each pain-wealth level; all you can do is compare and contrast one pain-wealth level with another. This means that there is no single, privileged stance from which you can pick out a 'best' outcome. And if your evaluations (or choice dispositions) are context-sensitive in this way, then it comes as no surprise that as the context of evaluation changes over the course of a sequence, you end up making inefficient choices. Because there is no single stance from which you can pick out a 'best' option, your views about what is preferable shift with changes in choice context, and there is no single, context-invariant strategy that you can stick to and arrive at an optimal outcome. If the world were structured more simply, containing a single kind of good or goods that can be traded off nicely, then there would be a single stance from which you could transitively order outcomes and pursue a unified, context-invariant strategy. But the world is not like that. So, it comes as no surprise that changes in context may induce inefficient decision-making (even if each individual decision is optimal in its context).

Rabinowicz (2014) argues that we should interpret the Money Pump pragmatically: *if* you want to avoid the costs of disunified decision-making (i.e., the costs of not being able to make all your decisions at one time, hence avoiding context-shifts), then you had better respect Transitivity. But Rabinowicz acknowledges that this is a conditional recommendation. While some may wish to avoid the costs of disunified decision-making, there is no categorical imperative to do so. And as someone who finds canonical cases of cyclic preferences compelling, I see no reason to avoid

[177] Levi (2002) thinks that if you engage in sophisticated choice, then you know that you will not be able to carry out any of the plans that save you a dollar. So, he argues, those plans are not feasible and you cannot be criticised for failing to implement them.

paying all costs of a disunified mind. After all, the Money Pump does not presuppose that I will (or am likely) to face a cunning exploiter; if I have a preference cycle cricket > golf > basketball > cricket but only ever play squash, then I do not expect to act on the basis of those cyclic preferences. Moreover, we all pay taxes for good reasons. If value is complex (being essentially comparative, multi-dimensional, etc.), then my considered judgements or deeply engrained tastes might mandate cyclic preferences—those considered judgements and engrained tastes just make me subject to certain taxes. So, to say that I am exploitable in the Money Pump is just to say that context matters in a certain way when I evaluate options; and if I really think that context so matters, I might accept the inefficiency.

An analogy with imprecise credences and ambiguity aversion is helpful here. Elga (2010) argues that imprecise credences make you vulnerable to exploitation in sequential choice problems, and it is well-known that ambiguity-aversion may lead you to select dominated sequences of options. In light of such cases Bradley (2017, p. 286), who defends imprecision and ambiguity-aversion, responds:

> 'There is no doubt that to have preferences that can be exploited in this way can be detrimental. If I could change them, that would be to my advantage, but in reality, transforming preferences can be expensive; often prohibitively so. In any case instability of preferences is not irrationality; nor, more generally, is vulnerability to exploitation a sure sign of it. So the possibility of exploitation does not in itself show that ambiguity aversion is irrational. The ambiguity averse agent's preferences impose costs on her that an ambiguity neutral agent does not face. But like agents with expensive tastes, ambiguity averse agents simply have to do the best they can with the preferences they are endowed with.' (Bradley 2017, p. 286)

The basic response in favour of cyclic preferences is the same. In the case of ambiguity-aversion, there is no single, precise probability function that guides your choices, so changes in information can induce predictable preference changes. You therefore vacillate in your judgements over time, leading you to be worse off than you would be if you could settle on a single, precise probability judgement with which to guide your choices. But no matter—the world is epistemically complex, and if you cannot settle on a single probability function that guides your choices, you might expect to pay taxes. And similarly with value. When outcomes cannot be transitively ordered, this means that there is no single, privileged way of aggregating all the things we care about that leads to a one-place utility function. There is no such thing as 'the utility' of an outcome. So, in the Money Pump you vacillate over the course of the sequence as you face different comparisons between options, sometimes favouring a and sometimes disfavouring it. But no matter, the world is complex, and if there is no single utility function that is robust across contexts, you might expect to pay taxes. Again, the *mere* fact that you choose dominated plans highlights the way in which your preferences are context-sensitive; but absent a good reason to think that preferences should be context-insensitive, I deny that such losses are problematic.

But perhaps you are not persuaded. After all, losing money is *bad.* And since you prefer to be better off, it is self-undermining to adopt preferences that make you worse off than you need be. McClennen (1988, p. 528) for example thinks that rationality constraints earn their keep only if violating them 'involves the agent in choice of means insufficient to his ends'. Cantwell (2003, p. 383) spells out an argument along these lines:

(i) *Preferences are incoherent if and only if there are situations where you can see that following those preferences is bad for you.*

(ii) *[The Money Pump] is sound just in case it shows that your preferences are incoherent.*

The idea here is that if *you* can see that cyclic preferences are bad, it is incoherent to stick with them.[178,179] For example, while you might adopt a rule for aggregating dimensions of value that results in cyclic preferences, if you can see that doing so bad for you, then you will see that you are better off not doing so. This argument is ingenious in part because it explains where normativity gets its force from—even someone like me can see that rational agents prefer to be better off and so, if coherent, will reject preferences they can see are bad for them.

We must take care, however, in claiming that some preferences are bad for you. The Money Pump shows that there is *a* situation in which following cyclic preferences leads to a sub-optimal outcome. But why conclude that your preferences are bad *all things considered* because there is *a single* situation in which those preferences result in a sub-optimal outcome? Cantwell (p. 383) claims that your preferences are bad 'if and only if *there are situations* where [you judge] it bad … to follow [your] preferences' (italics mine). But again, this is to move from a claim about how you assess your preferences locally (i.e., in some situations) to an all things considered evaluation of your preferences. And it is entirely coherent to judge some preferences to be bad locally but maintain that they do best given the totality of decision situations you might face. In principle

[178] For present purposes, I assume that if you can see that following your preferences is bad for you, there is some other set of preferences you could adopt (or act on the basis of) that would improve your situation. If not, then it might be coherent to stick with your 'bad' preferences simply because there is no alternative.

[179] Cantwell's own definition of what it takes for preferences to be 'bad' for you is more complicated than McClennen's. I first deal with the argument assuming that 'preferences that are bad for you' are 'preferences that foreseeably make you worse off than you need be'. Later, I deal with Cantwell's more subtle notion of what it takes for preferences to be bad for you, though the basic response remains the same.

then, it is coherent to stick with preferences that are bad in some situation(s), provided you think they are not bad globally.

I think you can maintain that exploitable cyclic preferences are not bad globally. Grant that you end up with a sub-optimal outcome in the Money Pump. What alternative do you have? You might move to the transitive ordering $a > b > c$ with $a > c$. But then consider how you would choose in the following:

Simple Choice: I offer you a choice between a and c.

Were you to adopt the transitive ordering above, you would choose a here. But you currently prefer $c > a$. So were you to choose a in the Simple Choice, you would make a choice that is mistaken by your current lights. Choosing a over c means choosing something that you currently think is worse for you.[180] You therefore recognise that were you to adopt the transitive ordering $a > b > c$, you would do something that is bad by your current lights in the Simple Choice.

Indeed, for any transitive ordering you might adopt, there will be some decision situation in which that ordering requires you to make a choice that you currently think is mistaken.[181] So, any transitive ordering you might adopt would require that in *some* situation you make a choice that leaves you worse off than you need be.[182] Therefore, though there exist situations in which following your cyclic preferences is bad for you (e.g., Cantwell's Money Pump), no transitive ordering would consistently leave you better off in every decision situation. To illustrate, recall my preference cycle $\text{Bach} > \text{Dylan} > \text{Beethoven} > \text{Dylan}$. I might adopt a transitive ordering to avoid possible inefficiency in a Cantwell-style case. But say that that transitive ordering requires me to choose Dylan over Bach. I currently balk at choosing Dylan over Bach—the genius of Bach *clearly* blows Dylan out of the water! Given that I might easily face a Simple

[180] Note that if we analyse preferences in terms of what is better for you or what you have more reason to choose, then by choosing a in the Simple Choice, you choose something that is worse for you or that you take to have reason not to choose. Likewise, if we analyse what is better for you in terms of preference satisfaction, then choosing a in the Simple Choice is worse for you than choosing c by your current lights.

[181] Proof: If a transitive ordering over $\{a, b, c\}$ contains at least one strict preference relation, then it contains at least one of $b > a$, $c > b$, or $a > c$. Since each of these disagrees with one of your current strict preferences, we can construct a pairwise choice between one of $\{a, b\}$, $\{b, c\}$, or $\{a, c\}$ in which you choose something that you currently strictly disprefer. If the transitive ordering contains no strict preference relations, then it is $a \sim b \sim c \sim a$. We can then use the Aversion to Small Fees assumption (which the Money Pump relies on): there is a quantity of money, $\$\epsilon$, such that $c > c - \$\epsilon$. Since $a \sim c$ in the transitive ordering, and assuming that $c > c - \$\epsilon$ in the transitive ordering, it follows that $a > c - \$\epsilon$ in the transitive ordering. So, you would choose a from $\{a, c - \$\epsilon\}$, which you currently strictly disprefer.

[182] Here, the friend of Transitivity might simply say that this is all the more reason to think that your preferences are bad—you are so sick that no medicine will cure you! Perhaps—but that will not convince the reasonable agent with cyclic preferences that *they* should judge a change in preferences to be all-things-considered desirable.

Choice between Dylan and Bach tomorrow, why would I commit to making an absurd choice—Dylan over Bach, *seriously!*—to save myself a dollar? Indeed, I might take the strength of my reasons to choose Bach over Dylan to be far greater than the strength of my reasons not to lose a dollar. I am better off by my current lights choosing Bach over Dylan in a pairwise choice, so I think it coherent to stick with my current preferences, even if I leave myself open to possible taxes.

In what sense then are cyclic preferences bad by my own lights, all things considered, if they result in inefficiency in the Money Pump? Pragmatic arguments are supposed to provide a non-circular reason why agents will not pay the cyclic preference tax. But given that non-exploitable preferences would lead me to make bad choices, I may judge the cyclic preference tax to be worth paying. Indeed, given that situations like Simple Choice are far simpler than the Money Pump, I may take myself to be more likely to encounter them than Money Pumps. So, I see that avoiding Money Pumps would require a substantive realignment of my ends, values, or goals. And saving a dollar need not be, by my lights, a decisive reason to undergo such a realignment.

Now, the discussion so far has rested on an intuitive definition of 'having preferences that are bad for you' as 'having preferences that lead to dominated outcomes'. Cantwell has a slightly more complicated analysis of what it takes to have preferences that are bad for you, but the argument so far applies equally to his analysis. His key idea is that a *bad plan* at decision node m is one that you do not judge at m to be worth following through. Slightly more precisely (p. 387), in a deterministic decision tree (i.e., one in which nature can make only a single move in response to each of your moves), a plan is bad at decision node m if (i) the plan involves some future choice at node n, (ii) following the plan from n leads to final outcome o, and (iii) $o \notin c_m(O_n)$, where O_n is the set of outcomes that you could receive on making some sequence of choices from n, and c_m is your choice function at node m.[183] Essentially, this says that a plan is bad at m if at m you judge the plan to be worth deviating from at some stage.

Though Cantwell's overall view contains various elements that do not bear directly on the current discussion, it entails (p. 388) that your *preferences* are incoherent (and hence bad) if at the start of some decision tree, every non-dominated plan is bad. So, *your preferences are bad if, in some decision situation, they tell you that all optimal plans are worth deviating from.*

[183] Recall that a choice function selects permissible options from option sets. In the next Chapter 6, Section 6.5, I dispute that O_n is the only set relative to which you should assess choices at n, which further problematises Cantwell's characterisation of bad plans.

Here is how cyclic preferences are bad on Cantwell's definition. In the Money Pump, say that at node 1 the optimal plans are those that yield any of $\{a, b, c\}$. Each such plan requires that you choose 'Up' at node 2, but at node 2 you face a choice between $\{a, c - \$1\}$, $\{b, a - \$1\}$, or $\{c, b - \$1\}$, so going 'Up' at node 2 means making a choice that you currently think is impermissible (e.g., the plan that yields a requires you to choose a from $\{a, c - \$1\}$, though you currently think $c - \$1$ is the choiceworthy option in this set). So, each optimal plan in the Money Pump is one that you think you should deviate from at some stage. This means that your cyclic preferences are incoherent, and you can see that they are bad.

The basic response to Cantwell's argument should here be clear. Firstly, context-sensitivity explains why optimal plans might not be worth carrying through. As you progress through the course of a decision tree, your choice context changes and the stance from which you evaluate options changes; so, though you might wish you could stick to an optimal course of action, the fact that you cannot do so merely reflects the way that context matters to you. That in and of itself is not inconsistency but context-sensitivity. Secondly, your preferences are not all things considered bad merely because they render optimal plans not worth following through *in some decision situation*. You might look at the optimal plans in the Money Pump and wish you could follow them. But then you might look at situations like the Simple Choice and conclude that a transitive ordering would require you to behave foolishly in that situation. So, you might agree with Cantwell that your preferences are bad locally (i.e., in the Money Pump) because they render optimal plans infeasible. But you might coherently stick with those locally bad preferences because they allow you to make the right choices globally—deviating requires you to make some blunder like choosing Dylan over Bach. Though the Money Pump shows that there is a local mismatch between feasible plans (or those that you think it worth following through) and optimal plans, it is coherent to stick with those preferences despite this mismatch.

Now, I have focussed on the fact that transitive orderings require you to make choices that are mistaken by the lights of your current cyclic ordering. The defender of the Money Pump might respond that what matters are your *ex post* views after changing your preferences. Though in the Simple Choice you currently judge it a mistake to choose a over c, you would not judge this to be a mistake on adopting the transitive preferences that recommend that choice. So, while cyclic preferences lead to an outcome that is sub-optimal by your current lights, no transitive ordering leads to an outcome that is sub-optimal by its lights. And perhaps this gives you decisive reason to adopt transitive preferences?

This response would show too much. I am going to the dreaded dentist this afternoon, and there is nothing I can do about that fact. The fact that I am better off preferring to go to the dentist by the lights of someone with those preferences does not mean that *I* have decisive reason to adopt those preferences. I should act by the lights of the preferences I currently have, not the preferences of some merely possible counterpart. Instrumental rationality is about specifying appropriate means to your ends, regardless of what those ends are. To say that you should adopt ends such that you maximise utility given your circumstances gets things the wrong way around—doing so places substantive constraints on your desires, tastes, or goals such that you end up maximising utility. The correct picture of decision theory, rather, is one on which you maximise instrumental utility in light of your desires, tastes, or goals. So, it would be a mistake to judge my current preferences by the *ex post* views I would have on changing my preferences. The Money Pump is compelling, if anything, because it highlights a problem that I can see with my own preferences. So the fact that my transitive counterpart is happy by their lights does not change the fact that *I* think it a mistake to behave as they do in the Simple Choice.

5.2.2 The Synchronic Money Pump

Gustafsson (2013) has proposed a *Synchronic Money Pump*, which is closely related to an argument due to Levi (2002). The case is simple: if I have intransitive preferences $a > b > c > a$, then simply offer me a choice from the set $\{a, b, c\}$ and I am guaranteed a loss since, whatever I choose, I choose something that is dispreferred to an available alternative. More precisely, recall that cyclic preferences require that you make an irrational choice given Davidson et al.'s (1955, p. 145):

> **Non-Dominated Choice Principle:** $f \in \boldsymbol{c}(O)$ only if there is no $g \in O$ such that $g \succ f$.

In response, I deny that choosing, say, a from $\{a, b, c\}$ is *dominated* in any normatively significant sense. Dominated options are those that are in *no sense* as good as an available alternative. Dominance principles are normatively compelling precisely because dominated options are unambiguously worse than some available alternative, so it would be self-undermining to choose them. If Gustafsson's synchronic Money Pump really did show that agents with cyclic preferences must choose dominated options, then they would recognise that their preferences are self-undermining because they force them to make choices they themselves can see are mistaken.

But, given that $a > b > c > a$, we can easily rationalise choosing at least one option from $\{a, b, c\}$. Say that you choose a. True, by choosing a you pass up the opportunity to have c, which you prefer. But there is something that can be said for a that cannot be said for c: a is preferred to b, whereas c is dispreferred to b. So, it is false that a is worse than c in every sense when b is available. The agent with cyclic preferences looks at the set $\{a, b, c\}$ and sees something in favour of each option (each option is preferred to another) and also sees something against each option (each option is dispreferred to another). If that is all the structure we have, then no option is worse than another in every sense, and there is no clear sense in which any choice is mistaken. At the very least, there is no reason for the agent with cyclic preferences to look at the set $\{a, b, c\}$ and think that their preferences are self-undermining.

This is again where context of evaluation matters for the agent with cyclic preferences. If $a > b$, then choosing b from $\{a, b\}$ is irrational. But, as has been pointed out by many (e.g., Anand 1993, Rabinowicz 2000), the fact that b is impermissible in $\{a, b\}$ does not mean that it is impermissible in $\{a, b, c\}$. The absence of c in the former means that you have decisive reason not to choose b; but the presence of c in the latter means that you have reasons against a, so b may now be your best bet. It would be smuggling in reasoning from the transitive setting to assume that because b is impermissible in $\{a, b\}$, this somehow means that b is ranked 'below' a in an absolute or context-free way. And so we have no reason to think that b is impermissible whenever a is—a local evaluation of impermissibility does not entail a global prohibition on b. Just as a squash player who loses a single match might still win the entire tournament, a strictly dispreferred option might be choiceworthy because of global properties of an option set (i.e., considering the totality of comparisons you might make between options). And again, there is nothing self-undermining in choosing an option that is beaten locally by some available alternative, provided that the option does best globally.

The lesson is that given cyclic preferences, the correct standard of rationality is not maximal preference satisfaction—one that requires you to choose an option to which none is strictly preferred. If your attitudes or desires are essentially comparative, then an option set need not contain an option that wins every pairwise comparison. But again, that is just what we might expect in a complex world with multiple dimensions of value, context-sensitive evaluations, and so on. The Non-Dominated Choice constraint is therefore overly demanding.

5.3 A Risky Challenge[184]

Standard Money Pumps do not undermine the case for rational cyclic preferences. While agents with cyclic preferences might wish that they were not exploitable, they need not realign their goals to avoid exploitation. It is therefore coherent to stick with exploitable preferences. So far, following much of the philosophical literature, I have focussed on intransitivity in the riskless context. I now discuss the challenge that risky cases pose for cyclic preferences.

To orient the reader, let me state my overall conclusion regarding the risky case I discuss: I have no firm conclusions. While broadly sympathetic to cyclic preferences, I feel troubled by the tensions that risky cases generate for agents with such preferences. And we will see that risky cases force us to reject at least one principle that seems close to decision-theoretic bedrock—I am unsure how to resolve debates that hinge on issues so close to bedrock. So, you might take this section as a map from an area underexplored by philosophers (cyclic preferences) to an ongoing debate (what the foundational concepts and commitment in our decision theory should be). The upshot will either be a novel counterexample to cyclic preferences *or* a choice point that helps us hammer out the shape of decision theory with cyclic preferences.

Take three outcomes, o_1, o_2, o_3, such that $o_1 > o_2 > o_3 > o_1$. Again assume Availability of Small Fees: there is some small fee that you pay to satisfy your preferences but that you absolutely prefer not to lose (again, I will call this fee a dollar, but you could make it as small as you like).

Now consider the following case:

> *The Easy Lottery*: Yesterday, a ticket was drawn from a fair 3-ticket lottery. Ticket I yields o_1, Ticket II yields o_2, and Ticket III yields o_3. Before I reveal which ticket was drawn, you can pay a dollar to *permute* the outcomes, such that Ticket I now yields o_3, Ticket II now yields o_1, and Ticket III now yields o_2.[185] That is:

	I	II	III

[184] The kind of case I discuss has been discussed by some economists. See, for example, Loomes and Sugden (1982, p. 822), Fishburn (1988, pp. 188-189), and Quiggin (1990, p. 508). Many such discussions are descriptive—which is appropriate given the work that economists are doing—rather than normative. Loomes and Sugden assume that cyclicity is due to *ex post* feelings of regret or rejoicing, which influences their discussion of risky cases significantly. Since I adopt an approach on which outcomes capture *all* relevant features, including regret and rejoicing, my discussion proceeds along different lines to theirs.

[185] If you are not prepared to pay a dollar to satisfy each of your pairwise preferences, substitute this for any small fee you like.

Do Nothing	o_1	o_2	o_3
Pay to Permute	$o_3 - \$1$	$o_1 - \$1$	$o_2 - \$1$

What should you do in the Easy Lottery? On the one hand, you surely cannot be required to Pay to Permute—paying to relabel the tickets in a fair lottery is just a waste of a dollar. On the other hand, you seem to be required to Pay to Permute—whichever ticket was drawn, you strictly prefer that you do so. So, whichever verdict we give, it seems that you violate some plausible rational constraint. Perhaps this gives us reason to reject the rationality of the cyclic preferences that give rise to the puzzle? Or perhaps accepting the rationality of cyclic preferences forces us to give up on some highly intuitive rational constraint? I formalise this tension before exploring it in more detail.

5.3.1 The Relabelling Principle

Most people I have talked to say that you ought to Do Nothing in the Easy Lottery. Indeed, if rationality requires that you relabel the tickets of a fair lottery, then you might think we have lost our grip on what rationality means. The idea behind this (overwhelmingly strong) intuition is that relabelling changes nothing of importance, so sacrificing a dollar to relabel is in no sense an improvement.

Formally then, say that g is a relabelling of f if there is a probability-preserving bijection $\rho: S \rightarrow S$ such that for each $s \in S$, $g(\rho(s)) = f(s)$. An intuitive principle is then:

> **The Relabelling Principle:** If g is a relabelling of f, then $\neg(g \succ f)$.

On the assumption that you are always required to pay some small fee to satisfy your preferences, this amounts to the principle:

> **The Relabelling Principle*:** If g is a relabelling of f, then you are not required to pay a fee to trade g for f.

If you (i) leave an act's outcomes unchanged, and (ii) shift those outcomes to states of the same probability, then you do not make the act *better*. At the very least, you cannot be *required* to pay to merely re-arrange the states that outcomes occur in. And since Pay to Permute is just paying a fee to have a relabelling of Do Nothing, the Relabelling Principle says that you are *not* required to Pay to Permute in the Easy Lottery

The Relabelling Principle is extremely weak. Meacham (2020, p. 1000) takes Stochastic Equivalence (that when two outcomes induce the same probability distribution over outcomes, one is permissible if and only if the other is) to be 'like a Moorean fact'. My intuition goes the same way as Meacham's here. And if you are prepared to pay a small fee to satisfy your preferences, the intuition becomes even stronger—if anything is self-evident about rationality, it is that you do not make an act *better* by subtracting a dollar from its outcomes while leaving the probabilities of outcomes unchanged. It is hard to give a principled argument for the Relabelling Principle; nonetheless, I simply *cannot* be rationally required to prefer to Pay to Permute in the Easy Lottery. I discuss the role of this intuition in more detail later. For now, note that insofar as the Relabelling Principle captures a plausible principle of rationality, you are permitted *not* to Pay to Permute.[186]

5.3.2 State-wise Dominance

If you Pay to Permute, then you get something that you strictly prefer, regardless of how the world turns out to be. And State-wise Dominance says that if one act yields a strictly preferred outcome however the world turns out to be, then you strictly prefer that act *simpliciter*. Hence, you strictly prefer Pay to Permute in Easy Lottery.

Note that the subtleties that arose for State-wise Dominance in Newcomb's Paradox do not arise here—we can let the states (i.e., which ticket was drawn) be causally, counterfactually (on standard or backtracking readings), probabilistically, or whatever you like, independent of your choice. The paradox generated by the Easy Lottery does not rely on subtle questions about what to hold fixed when deliberating about your acts. It arises instead because of the structure of cyclic preferences themselves—cyclicity means that each outcome is dispreferred to another in the cycle, so we can construct cases in which a seemingly preferable act is dispreferred in every state.

[186] Note that the Relabelling Principle is closely related to First-Order Stochastic Dominance (First-Order Stochastic Dominance trivially entails the Relabelling Principle). Why not say that you ought not Pay to Permute on the grounds that doing so violates First-Order Stochastic Dominance? Answer: First-Order Stochastic Dominance itself yields highly counterintuitive verdicts given cyclic preferences (for an example, see Fishburn 1978, pp. 1270-1). Nonetheless since the Relabelling Principle is entailed by First-Order Stochastic Dominance given Transitivity, we might think of the Relabelling Principle as a minimal extension of First-Order Stochastic Dominance to the intransitive setting. In that sense the Relabelling Principle is in the family of dominance principles.

Once suitably qualified, most philosophers agree that State-wise Dominance is a non-negotiable constraint. Not only does Savage's EU respect it, risk-sensitive generalisations of Savage's theory such as Risk-Weighted Expected Utility Theory (Buchak, 2013) and Grant et al.'s (2000) Decomposability-satisfying theory respect it. Insofar as State-wise Dominance captures a highly intuitive principle of rationality, you are required to Pay to Permute in the Easy Lottery.

Worth noting here is that in cases of incomplete preferences, many reject the principle:

> **Negative State-wise Dominance:** If for each state s, $\neg(f(s) \succ g(s))$, then $\neg(f \succ g)$.

This is motivated by considering a well-known case from Hare (2010), which I discussed in the Introduction. Say that you have a preference gap r and s, and consider the following gamble on the toss of a fair coin:

	Coin Heads	Coin Tails
Unsweetened Gamble	r	s
Sweetened Gamble	$s + \$1$	$r + \$1$

Though you do not strictly prefer the Sweetened Gamble in each state, many think you should prefer the Sweetened Gamble. This means that Negative State-wise Dominance is false (so argue Hare 2010, Bader 2018, and Wilkinson Forthcoming).

I emphasise that it is both coherent and natural to reject Negative State-wise Dominance (in cases involving preference gaps) while upholding State-wise Dominance itself. For example, you might reject Negative State-wise Dominance because it conflicts with plausible requirements on how incomplete preferences cohere with possible completions of those preferences—for example, that your incomplete preferences should be compatible with at least one transitive completion of those preferences (cf. Hare 2010, p. 243). But if, as most do, you take State-wise Dominance to be a constraint on coherent, complete preferences, then this justification for rejecting Negative State-wise Dominance *requires* us to endorse State-wise Dominance. In this spirit Rabinowicz (Forthcoming) has recently argued against Negative State-wise Dominance (which he calls Complementary Dominance) using State-wise Dominance as a premise. So, though many reject Negative State-wise Dominance, this does not mean that all plausible forms of state-wise reasoning should go. Indeed, following Rabinowicz, we might reject Negative State-wise Dominance out of a desire to uphold State-wise Dominance itself.

5.4 Counterexample or Choice Point?

If we accept the Relabelling Principle and State-wise Dominance, then cyclic preferences must go. More precisely, the argument is:

1. The Easy Lottery is a possible decision situation.[187]
2. The Relabelling Principle holds, so a rational agent with cyclic preferences is not required to Pay to Permute in the Easy Lottery.
3. State-wise Dominance holds, so a rational agent with cyclic preferences is required to Pay to Permute in the Easy Lottery.
4. Contradiction (by 2 and 3).

Therefore,

5. The agent with cyclic preferences is not rational.

If sound, this argument fills an important gap. I have already rejected the standard, diachronic Money Pump for acyclicity. And I am not the only one suspicious of diachronic arguments—sophisticated choice strategies, backwards-looking strategies, and a move towards time-slice rationality have convinced many that standard Money Pumps fail. The Easy Lottery, however, is synchronic and so does not run into these problems. Moreover, unlike Gustafsson's Synchronic Money Pump, the Easy Lottery does not appeal to Davidson et al.'s Non-Dominated Choice constraint, which is arguably overly demanding. So, the Easy Lottery represents a novel

[187] I include this premise because we might accept rational dilemmas if they only arise in decision situations you can never face. A referee raises the following argument against this premise: If we individuate outcomes based on modal properties (i.e., what else you could have got in some decision), it may not be possible to embed o_1, o_2, o_3 in the Easy Lottery as I have supposed. For example, you might prefer $o_1 > o_2$ but deny these outcomes can appear as outcomes in the Easy Lottery (since the Easy Lottery *could* yield o_3, and 'getting o_1' and 'getting o_1 as the result of a lottery that could yield o_3' differ with respect to modal properties). So, though $o_1 > o_2 > o_3 > o_1$, this cycle may only hold for those outcomes *qua* riskless outcomes, making the Easy Lottery a case that you can never practically face. I reject this response for two reasons. (i) Even if you are permitted to care about modal features, it is unclear that you *must.* An agent with cyclic preferences can reasonably deny that they care about modal features of outcomes in the way that this response requires; such an agent treats the outcomes in the Easy Lottery as the very same outcomes that appear in the initial preference cycle. Again, since we are not in the business of dictating tastes, we should allow such modally-insensitive preferences, meaning the Easy Lottery is possible. (ii) The entire decision-theoretic project is about assessing risky acts based on preferences over riskless outcomes. To deny that riskless outcomes can consistently be embedded in risky lotteries is to give up on this project. For present purposes then, I work with the standard decision-theoretic assumption that the outcomes you have preferences over in the riskless context are the same outcomes that result from risky gambles, though they might differ in their modal features.

counterexample to cyclic preferences with significant advantages over existing alleged counterexamples.

Alternatively, we might take the Easy Lottery not to be a counterexample, but to highlight a *choice point* for the defender of cyclicity. Though the Relabelling Principle and State-wise Dominance are both plausible given Transitivity, cyclic preferences force us to rethink at least one fundamental commitment of rationality. If so, the Easy Lottery presents a *challenge* for the defender of cyclic preferences to get clear on which principle we should reject and to tell a story about why we should reject it. My goal in the remainder of this chapter is not to decide the upshot of the Easy Lottery—I will not myself settle whether it constitutes a counterexample or choice point, or how to respond to that choice point. Instead, I will highlight some features of (and further challenges raised by) each possible response to the case. We will also see that with the challenge comes an opportunity to spell out the details of a normatively sound decision theory without Transitivity.

5.4.1 Denying the Relabelling Principle

The first possible response is to deny the Relabelling Principle—rationality simply can require that you prefer to relabel the tickets of a fair lottery, as counterintuitive as that might seem.

This response is perhaps coherent. After all, when deciding between the Relabelling Principle and State-wise Dominance we are approaching decision-theoretic bedrock, and it is hard to know what counts (or whether anything could count) as a decisive argument for one principle over the other. So, the defender of State-wise Dominance might simply thump the table and insist that, despite appearances, you must Pay to Permute.

I am unsure how persuasive the table thumping approach is. It strikes me as carrying a significant methodological cost. We should be able to offer some account of why reasonable agents ought to behave rationally, and if we deny the Relabelling Principle, then it is unclear where the normative force of 'ought' comes from. What could you say to someone like me who finds it deeply implausible that they are required to Pay to Permute? I take myself to save a dollar by refusing to Pay to Permute, and I want to save a dollar. So if you insist that I am irrational, then I want some story about why I should care about rationality in your sense rather than mine. Without such a story, State-wise Dominance (along with cyclic preferences) makes murky the connection between the outputs of decision theory and what we actually take to be the best means to our ends.

You might point out that I disagree with my fully informed views about which option is preferable. But I respond that it is coherent to disagree with my fully informed self in this case, since my fully informed self does not face the uncertainty that I face. From my current epistemic perspective, Do Nothing might yield any of o_1, o_2, or o_3, while Pay to Permute might just as easily yield any of $o_1 - \$1, o_2 - \$1, o_3 - \$1$. Following Bader (2018, p. 501) then, I might think that the focus of instrumental rationality is on how each act might turn out (epistemically speaking) rather than how each act *will* actually turn out. Therefore, though I will actually receive a preferable outcome on Paying to Permute, that fact is irrelevant (from my *ex ante* perspective) since for any outcome that Pay to Permute might yield, Do Nothing might just as easily yield that outcome plus a dollar.[188] In this way, the uncertainty I currently face, and that my fully informed self does not face, explains why I coherently disagree with my fully informed self about the best means to my ends.[189, 190]

Of course, we typically think that an act that does better in each state is foreseeably better for us. But cyclic preferences drive a wedge between our fully informed (or future) views and our current views. Currently, I could end up with any of o_1, o_2, or o_3 with equal probabilities, so I do

[188] Moreover, recall that Ahmed and Price 2012 take repeated trials of the same situation to illustrate what does foreseeably in that decision. We can give such an argument for Do Nothing over Pay to Permute. In N trials of the Easy Lottery, for large enough N, the expected return of Do Nothing is receiving $\frac{N}{3}$ of each of o_1, o_2 and o_3. The expected payoff of Pay to Permute is identical, except that you lose $\$N$. So, assuming that more money is absolutely preferred to less, the Do Nothing strategy has foreseeably preferable consequences.

[189] There is a loose parallel here with the debate between one- and two-boxers. Two-boxers like Joyce think that the two-boxer does as well as they could have done given how facts outside their influence are. But recall that Ahmed responds:

> '[N]either Irene nor any other agent cares, when choosing, about whether she does *better* than anyone, actual or possible, counterpart or not. Nor does she care about whether she will consequently regret her choice. She only cares about her terminal wealth. So Rachel's point is irrelevant [i.e., the point that she, as a two-boxer, did as well as she could holding fixed the predictor's prediction]. … What matters is only that she [Irene, the one-boxer] is *richer*.' (Ahmed 2014b, pp. 185-186)

We might respond similarly in the Easy Lottery: 'Nobody cares about whether they will consequently regret their choice, nor whether they do better than anyone, actual or possible, counterpart or not.' Instead, what matters is which act is foreseeably best, and there is a clear sense in which Do Nothing is foreseeably best—it does best on average—even if you will come to envy your counterpart who did otherwise.

[190] Loomes and Sugden (1982, see also Fishburn 1991, p. 119) claim that you might be required to Pay to Permute to avoid feelings of regret. This might be right insofar as Loomes and Sugden's model is a descriptive one that allows outcomes to be partially, rather than, fully described. But if outcomes fully describe your ends—as I assume here—then there can be no question of your paying to avoid some negative feeling (e.g., regret) in the Easy Lottery, since (dis)valuable psychological components of outcomes are already built into outcome-descriptions. Fishburn (1991, p. 119) further suggests that you might pay to relabel simply because of 'the importance of … being better off in the end'. But that begs the question—it assumes that it *is* important to bring about a preferable outcome *ex post* rather than, say, do what *ex ante* best promotes your goals.

not think it an improvement to perform an act that could yield one of $o_1 - \$1, o_2 - \$1$, or $o_3 - \$1$ with equal probabilities. What is *ex ante* best is not what is *ex post* preferable.

Now, the position I have sketched might be resisted—*you* might think of what is *ex ante* best differently, or you might have different intuitions about the case. Nonetheless, the challenge still stands: what can you say to someone who disagrees with you (or has different intuitions) to convince them that they should care about rationality in your sense rather than theirs? Given that rationality is supposed to guide and constrain our behaviour, there are some verdicts that strike me as so absurd that we have every reason to reject them in our normative theorising. At the very least, anyone who rejects the Relabelling Principle owes us a story as to where rationality gets its normative force from.[191]

5.4.2 Denying State-wise Dominance

I now turn to the second way of meeting the challenge posed by the Easy Lottery—reject State-wise Dominance. Clearly, the key advantage of this response is that it allows us to say the sensible thing in Easy Lottery. But virtually every normative decision theory respects State-wise Dominance—many take it to be a non-negotiable constraint (e.g., Buchak 2013, p. 37). Given how widely accepted State-wise Dominance is, this will strike some as a significant bullet to bite. Can we say anything to sweeten it?

We might motivate this position by noting that several philosophers have recently defended *prospectism*, the view that an option's choiceworthiness supervenes on the probability distribution it induces over outcomes. Though interest in prospectism is recent, it has been discussed under the guise of *the reduction principle* (see Fishburn 1988, p. 27) and, arguably, is implicit in von Neumann and Morgenstern's (1953) theory.

Formally, define an option $f = \alpha_1 o_1 + \cdots + \alpha_n o_n$ prospect as $P(f) = \{(o_1, \alpha_i), \ldots, (o_n, \alpha_n)\}$.[192] That is, f's prospect fully describes how likely f makes each outcome,

[191] Gilboa (2010) analyses a 'normative principle' as one that a decision-maker is motivated to adhere to once that principle has been explained to them. Pay to Permute really does seem absurd to me, and I deny that reasonable decision-makers will be motivated to Pay to Permute when State-wise Dominance is explained to them. So, at least on Gilboa's account of normativity, I am suspicious that we can motivate State-wise Dominance over the Relabelling Principle.

[192] Recall that, in the framework I adopt for this thesis, α_i is your credence in the states s such that $f(s) = o$. Since we could define states as probabilistically or causally act-independent propositions, prospectism is formally neutral between Causal and Evidential Decision Theory.

and nothing more. Prospectism then says that preferences supervene on prospects—there are no normative differences between acts with identical prospects. Do Nothing's prospect is:

$$P(\text{Do Nothing}) = \{\left(\frac{1}{3}, o_1\right), \left(\frac{1}{3}, o_2\right), \left(\frac{1}{3}, o_3\right)\}$$

And Pay to Permute's prospect is:

$$P(\text{Pay}) = \{\left(\frac{1}{3}, o_1 - \$1\right), \left(\frac{1}{3}, o_2 - \$1\right), \left(\frac{1}{3}, o_3 - \$1\right)\}$$

Now it is easy to see that any minimally plausible prospectist theory says that you ought to Do Nothing. Consider the following decision situation:

	I	II	III
Do Nothing	o_1	o_2	o_3
Pay to Permute*	$o_1 - \$1$	$o_2 - \$1$	$o_3 - \$1$

Everybody agrees that you ought to Do Nothing here (Pay to Permute* is just sacrificing a dollar for nothing). But Pay to Permute* and Pay to Permute have identical prospects. So, if you prefer Pay to Permute over Do Nothing but prefer Do Nothing over Pay to Permute*, then preferences supervene on something other than prospects. So, the prospectist must say that you prefer Do Nothing in the original Easy Lottery, which means rejecting State-wise Dominance. (Alternatively, the prospectist could deny that State-wise Dominance is well formulated in the first place. A decision theory that only mentions acts and probability distributions over outcomes might not include a well-defined notion of a state, in which case prospectists do not reject State-wise Dominance so much as deny that it is well-defined in the first place. Since I assume we can simply define states as a certain kind of act-independent proposition, I will talk as if the prospectist rejects State-wise Dominance.)

The prospectist position is intuitively plausible. Can we therefore simply appeal to prospectism and reject State-wise Dominance? Doing so has one revisionary implication that you might take to be a cost, which I now outline.

Begin with the plausible idea that the goal of decision theory is promote value, where 'value' here is understood specifically in terms of the kind of value constituted by preference satisfaction. If so, then it is incoherent to think that the best means to your ends *actually* violate your preferences over outcomes—there is no sense in which something can bring about certainly worse ends while being the best means to your ends. Ahmed and Spencer (2020, p. 1166) introduce a principle that connects objective-value facts to subjective oughts:

> **Bridge**: If options have objective values, then: if an agent's options are $a_1, \dots, a_n$ and the agent knows for certain that option a_i uniquely maximises objective value, then option a_i uniquely maximises subjective value.

Bridge is formulated in a context where instrumental utilities have an ordinal structure. This fails to hold if preferences are cyclic. Nonetheless, even if options might not have objective *one-place* values we can, in the spirit of cyclicity, accept that there are facts about which options are objectively more valuable than others—we move from (one-place) *objective values* to (two-place) *objective comparative values*. So, we might tweak Bridge to get a principle that is at least as plausible:

> **Bridge***: If options have objective comparative values then: if an agent's options are $a_1, \dots, a_n$ and the agent knows for certain that option a_i uniquely maximises objective value, then option a_i uniquely maximises subjective value.[193]

This strikes me as uncontroversial: if there are facts about which options are objectively more valuable than others, then options that are objectively more valuable are better means to your ends. And yet any minimally plausible account of objective value will say that Pay to Permute maximises objective value in the Easy Lottery. *As a matter of fact*, Pay to Permute brings about a preferable, so better for you, outcome. And again, it is incoherent to think that instrumental rationality aims to promote value while saying that an actually less valuable act is a better means to your ends. Again to draw on Schoenfield (2014), that would be instrumental value fetishism.

More generally, say that there are facts about what is objectively more valuable than what. Then such facts determine what you objectively ought to do. Jackson (1991, p. 466), for example, analyses what you objectively ought to do as the thing that as a matter of fact brings about the best consequences. Now, decision theory is not about working out what the objective oughts are. Rather, decision theory is about issuing subjective oughts. Nonetheless, it is implausible that there is no connection between the two. For example, an agent who said, 'I know that I objectively ought to do f, but I still think I needn't do f' would seem to be deeply confused

[193] Note that once we move to objective comparative values, we can say that a_i maximises value in $\{a_1, \dots, a_n\}$ if a_i is at least as valuable as every element of $\{a_1, \dots, a_n\}$.

(indeed, we might question whether they have really grasped what the word 'ought' means). So, just as with Bridge, we ought to insist on:

> **Deference to Objective Oughts (DOO):** If you know that you objectively ought to perform option f, then you subjectively ought to perform option f.[194]

And again, it seems that the sensible verdict in the Easy Lottery violates this principle. As a matter of fact, you ought to Pay to Permute—as a matter of fact doing so brings about better consequences. Therefore, subjectively you ought to Pay to Permute.

So, minimal constraints on the relationship between means and ends say that you should Pay to Permute in the Easy Lottery. The lesson is clear: if we accept the intuitive verdict in the Easy Lottery, namely that you are permitted to Do Nothing, then the goal of rationality cannot be to promote some objective quantity (utility, value, etc.). There are no facts about what is objectively better for you, what is objectively preferable for you, or what objectively you have most reason to do. If there were such facts, they would constrain your subjective evaluations or choice dispositions and require that you Pay to Permute in the Easy Lottery. So, the intuitive verdict in the Easy Lottery requires us to accept a radical view, one on which there are no utility- or value-facts that go beyond *ex ante*, subjective evaluations of options.

5.4.3 Moving Forward: Pluralism or Reflective Equilibrium?

What is the correct response to the Easy Lottery? I am genuinely ambivalent. One day, I think that Do Nothing is *obviously* the best means to your ends in the Easy Lottery. And since we are not in the business of dictating tastes, the permissibility of cyclic preferences means that State-wise Dominance should go. The next day, however, I think that there are *obviously* facts about which means objectively promote your ends, and reasonable agents should recognise that it is self-undermining to select means that lead to objectively worse ends. So, State-wise Dominance should stay; and since the Pay to Permute verdict is wildly implausible, I can see that my cyclic preferences over riskless outcomes do not cohere with my reasoned judgements in the risky setting. So, I am dragged into rejecting my cyclic preferences. But the next day the Humean in me wakes up, thinks that of course my cyclic preferences are reasonable, and so thinks that recognising objective value-facts pushes me to accept a radically counterintuitive verdict in the

[194] This does not say that in cases where the objective ought is known, the subjective ought is irrelevant or plays no action-guiding role. It simply says that the subjective ought should *agree* with the objective ought. DOO is therefore compatible with the thought that it is the subjective ought, and only the subjective ought, that is action-guiding.

Easy Lottery. It seems that I prefer to uphold the Relabelling Principle over State-wise Dominance, that I prefer to uphold State-wise Dominance over the rationality of cyclic preferences, but that I prefer to uphold the rationality of cyclic preferences over the Relabelling Principle.

I am unsure how to resolve this ambivalence. Perhaps your intuitions go differently to mine in the Easy Lottery and you have an easier time accepting the Pay to Permute verdict. Or perhaps you already accept a normative view on which there are no objective value-facts, and so Do Nothing is unproblematic. But for those who like me feel torn, or unsure how to persuade somebody who disagrees with them, I consider on possible routes forward.

The first route is to adopt a pluralist response (as do Horgan 1981 and Bales 2018 in response to Newcomb's Paradox). You might think that there are two legitimate ways of spelling out 'rationality' given cyclic preferences. On the one hand, there is the *ex post* standard of rationality, which takes preferences over options to be constrained by your preferences over actual consequences. On the other hand, there is the *ex ante* standard of rationality, which takes preferences over options to supervene on probability distributions over outcomes. And on still yet another hand, there is the transitive standard of rationality, on which we accept that both *ex ante* and *ex post* judgements matter and insist that they coincide. Just as with Newcomb's Paradox then, you might think that we are free to pick a standard of rationality that fits our intuitions. An advantage of this approach is that it recognises how close to bedrock some of the principles and intuitions we are dealing with are, and it allows that there might be no single, privileged way of trading between them.

A different response would be to look to a broader process of reflective equilibrium to decide between positions in the Easy Lottery. Though it is hard to give a decisive argument for or against, say, State-wise Dominance, you might think that a process of fitting intuitions to principles and vice-versa will eventually demonstrate that one standard of rationality is the best way of balancing our considered judgements.

As an illustration, here is one way that process might go. You might initially think that the debate between one- and two-boxers stalemated (as I have indeed argued in previous chapters!). But then you might note some parallels between the discussion here and that in Newcomb's Paradox. In both cases, it seems like a *prima facie* reading of State-wise Dominance leads to strange verdicts. Initially, as I did in Chapters 2-4, you might try to re-work State-wise Dominance. But then you might note that the Easy Lottery puts pressure not just on the details but the relevance of state-wise comparisons entirely. So, you may conclude that State-wise

Dominance is not as compelling as you initially thought—you might start to doubt it once we step outside of familiar settings. You might then think about conflicts between *ex ante* and *ex post* judgements more generally, say in the context of risk-sensitivity (e.g., Buchak 2013, Chapter 6.4) or population ethics (e.g., Nebel 2020, Blessenohl Forthcoming, Gustafsson Forthcoming). Depending on how those debates go, you might think that there are just too many cases where what seems best *ex ante* conflicts with what is best *ex post*, and you might be pushed to a view of instrumental rationality on which the best means to your ends are those that are best given your current epistemic perspective, which is determined solely by looking at prospects. And so you might think that as part of a process of reflective equilibrium, as sacrosanct as it seemed, even State-wise Dominance should go.

Now, the above is just a sketch of one way the larger decision-theoretic project could go. Perhaps it will go very differently. Before moving on, I simply want to emphasise two methodological points: (i) We need to be careful about letting some principle (e.g., State-wise Dominance) act as decision-theoretic bedrock without thinking through the implications of that principle in full generality, and (ii) Often pluralism seems like the right response to some tension locally, but when we look for similarities between debates, we might find that certain culprits keep turning up, and we might take the combined weight of several possible counterexamples to decisively refute a principle. Again, I will not here reject State-wise Dominance. But I will state that it is far from obvious in the Easy Lottery, and we need to balance the virtues and vices of rejecting State-wise Dominance in the context of our decision theory as a whole.

5.5 Risk and Intransitivity: Toward a General Proposal

We have our risky challenge: reject cyclic preferences or reject some plausible rational constraint. I now discuss how our verdict in the Easy Lottery informs our decision theory more generally.

If we reject cyclic preferences, then the lesson from the Easy Lottery is simple: return to Expected Utility Theory, Risk-Weighted Expected Utility Theory, Weighted Linear Utility Theory, or whichever acyclic theory of instrumental rationality you like.

But what if we decide to uphold the rationality of cyclic preferences? Then we must do decision theory without appealing to *the* utility of an outcome or *the* instrumental utility of an act. Nonetheless, in the absence of a one-place utility function u, we can compare and contrast outcomes. Indeed, doing so is natural given that preference is a two-place relation—though there might not be a fact about how good outcome o is, we may still ask how strong your preference for o is over some other outcome. In this way we move from decision theory built on *utility* to one built on *utility-differences*. This is a familiar move. Easwaran (2014b), in perhaps the leading realist approach to decision theory, takes utility-differences as primitive in his theory. More generally, the invariant quantity in standard decision theory is not utility (which is only unique up to positive affine transformation) but ratios of utility-*differences*: for outcomes o_1, o_2, o_3, o_4, it is $\frac{u(o_1)-u(o_2)}{u(o_3)-u(o_4)}$ that is fixed independently of our choice of representation.[195] As further evidence that comparative or contrastive measures of utility are fundamental to decision theory, plenty of philosophers are already convinced that in situations where expected (or instrumental) utilities are undefined, we can still compare acts with a two-place function (see, for example, Bartha 2007, Colyvan 2008, and Meacham 2020). So, accepting cyclic preferences can be thought of as part of a general move towards a contrastive or comparative utility theory.[196]

I say that $\phi: O \times O \rightarrow \mathbb{R}$ is a utility-difference function if it satisfies (i) $\phi(o_1, o_2) \geq 0$ if and only if $o_1 \geq o_2$, and (ii) $\phi(o_1, o_2) = -\phi(o_2, o_1)$. The first condition says that utility-differences represent preferences, while the second says that such differences are order-invariant. (If strength of preference is not order invariant, then following Buchak 2013 p. 149, it is unclear that you count as having 'clear desires' in the first place.) I initially assume nothing more about utility-difference.[197]

Since we are interested in assessing means to ends, we need to say how utility-differences of outcomes constrain instrumental utility-differences of acts. Here we can note that different ways of calculating instrumental utility-difference will say different things about the Easy Lottery. So, the Easy Lottery helps us clarify the shape of instrumental rationality without Transitivity.

Firstly, take Fishburn's (1982) *Skew-Symmetric Bilinear Utility Theory* (SSB). Working in the von Neumann and Morgenstern framework (i.e., identifying each act with the probability distribution it induces over outcomes), Fishburn (1982) introduces the following preference axioms:[198]

[195] Zynda (2000), in the context of epistemology, claims that the 'real' features of your belief are those that are representation invariant. If we extend this idea to decision theory, then what is real are not facts about utility but ratios of utility-difference.

[196] This in turn is part of a move towards a contrastive or comparative view of normative philosophy more generally. For example, Temkin (2012) defends an essentially comparative view of value, and Snedegar (2017) defends a contrastive view of reasons.

[197] Others (e.g., Easwaran 2014b, p. 21) will impose stricter requirements on utility-difference that entail preferences are transitive.

[198] Fishburn's proof assumes a bounded outcome set (i.e., that there are outcomes m, n such that for all $o \in O, m \succcurlyeq o$ and $n \preccurlyeq o$). I take this to be a further structural assumption in what follows.

Continuity: If $f \succ g \succ h$, there exists some probability α such that $g \sim \alpha f + (1-\alpha)h$.

Mixture Dominance: If $f \succ g, f \succ h$, then for all $\alpha, f \succ \alpha g + (1-\alpha)h$. If $f \sim g$, $f \sim h$, then for all $\alpha, f \sim \alpha g + (1-\alpha)h$. If $f \prec g, f \prec h$, then for all $\alpha, f \prec \alpha g + (1-\alpha)h$.

Symmetry: If $f \succ g \succ h$ and $f \succ h$, then $\alpha f + (1-\alpha)h \sim \frac{1}{2}f + \frac{1}{2}g$ if and only if $(1-\alpha)f + \alpha h \sim \frac{1}{2}g + \frac{1}{2}h$.

He proves that if you satisfy these three axioms, then there exists a unique (up to positive scalar multiplication) utility-difference function $\phi: O \times O \rightarrow \mathbb{R}$ that represents your preferences over outcomes. Moreover, for acts $f = \alpha_1 o_1 + \dots + \alpha_n o_n$ and $g = \beta_1 o_1 + \dots + \beta_n o_n$, $f \succcurlyeq g$ if and only if $\Phi(f,g) \geq 0$, where Φ is defined:

$$\Phi(f,g) = \sum_{i,j} \alpha_i \cdot \beta_j \cdot \phi\,(o_i, o_j)$$

This rule for comparing acts is compatible with the Relabelling Principle and the verdict that you should Do Nothing in the Easy Lottery.[199]

Now, you might worry that SSB cannot be applied to cases like the Easy Lottery (thanks to an anonymous referee for raising this point). After all, Φ weights the utility-difference $\phi(o_i, o_j)$ by the probability of o_i under f multiplied by the probability of o_j under g. This seems to assume that f and g are independent lotteries. If so, you might think that SSB is only intended to apply when the outcomes of f and g are probabilistically independent. If so, then we cannot apply

[199] SSB satisfies the Relabelling Principle because it satisfies the reduction principle—it requires that you are *indifferent* between acts that induce the same probability distribution over outcomes. If ϕ is such that for all x, y, $\phi(x,y) > \phi(x, y - \$\epsilon)$ then SSB always says a prospect is made worse by subtracting $\$\epsilon$ from each outcome. To see this, note that if $f = \alpha_1 o_1 + \dots + \alpha_n o_n$ and $g = \alpha_1(o_1 - \$\epsilon) + \dots + \alpha_n(o_n - \$\epsilon)$, then:

$$\Phi(f,g) = \sum_{i,j} \alpha_i \cdot \alpha_j \cdot \phi\big(o_i, o_j - \$\epsilon\big)$$

$$= \sum_i \alpha_i^2 \cdot \phi(o_i, o_i - \$\epsilon) + \sum_i \sum_j \alpha_i \alpha_j [\phi\big(o_i, o_j - \$\epsilon\big) + \phi\big(o_j, o_i - \$\epsilon\big)]$$

Each $\phi(o_i, o_i - \$\epsilon) > 0$ by definition, so the first summand is greater than 0. And $\phi\big(o_i, o_j - \$\epsilon\big) + \phi\big(o_j, o_i - \$\epsilon\big) > \phi\big(o_i, o_j\big) + \phi\big(o_j, o_i\big) = 0$, so the second summand is greater than 0. So $\Phi(f,g) > 0$, as required. So, for appropriately structured ϕ, SSB always ranks a lottery over a stochastically equivalent lottery minus a fee.

SSB to the Easy Lottery, since in that case learning the outcome of one act tells us how the other act resolves.

But the discussion so far indicates another way of thinking about SSB, which does not presuppose that acts are independent lotteries. Formally, nothing in the SSB axioms (or von Neumann and Morgenstern's axiomatic approach more generally) specifies that acts are independent lotteries. And the motivation for Do Nothing in the Easy Lottery is that we need to consider each outcome that each act might bring about *ex ante*, regardless of whether those outcomes occur in the same state. Again, if we opt for the Do Nothing response to the Easy Lottery, then we follow Bader (2018, p. 501) in taking rationality to be a matter of evaluating acts based on what *might* happen, rather than what *will* or *would* happen as a matter of actual consequence. If states are defined, this will mean making inter-state comparisons—SSB does not restrict itself to comparing acts based on outcomes that can both occur given that some state holds. Instead, SSB essentially encodes the purely *ex ante* conception of rationality: it tells you to think about each outcome that f might bring about (weighted by its probability given f), each outcome that g might bring about (weighted by its probability given g) and consider the utility-differences of those outcomes. This is not to say that SSB relies on a convenient fictionalisation that acts are independent lotteries. Rather, because SSB focusses on probability distributions *ex ante*, it erases any distinction between dependent and independent lotteries when it comes to calculating instrumental utility-difference.

A further worry you might have about SSB is that it is an *expectational* model and so overly restrictive. Buchak (and I in Chapter 1) have argued that rational agents may care about more than the average payoff of a lottery—risk may matter such that the instrumental utility of a lottery is not its expected utility.

But SSB shows that while caring about expected *utility* may be overly restrictive, caring about expected *utility-difference* is a far more minimal and plausible position. Recall that Fishburn's axiomatisation of SSB relies on three axioms. The first, Continuity, is a structural assumption—it rules out outcomes like Pascal's Heaven and so avoids complications arising when some acts are incomparably better (or worse) than others. While perhaps not fully general, it is still interesting to consider the shape of instrumental rationality setting aside Pascalian outcomes. The second axiom Symmetry, in the presence of the other two axioms, plays an essential role in ensuring that utility-difference is order invariant (i.e., that $\phi(x,y) = -\phi(y,x)$ for all x,y).[200] And however

[200] See Section 6 of Fishburn (1982); Fishburn 1984, (p. 74) considers taking order-invariance (and hence axioms like Symmetry that underwrite it) as a 'definitional characteristic of strict preference'. Note that without Symmetry,

we think about utility-difference, if it is a normatively significant quantity, then its magnitude had better be order invariant. So, we can grant Continuity and Symmetry as structural assumptions. But then Mixture Dominance turns out to be *the* defining feature of expectationism about utility-difference! So, whether expectationism (about utility-difference) holds ultimately amounts to whether Mixture Dominance holds. And Mixture Dominance is extremely plausible. Like Betweenness, it requires insensitivity to pure randomisation—acts do not get better or worse simply by tossing a coin. At the very least then, we have a compelling answer to Buchak's question about what is so special about *expectations*: caring about expectations ultimately reduces to caring about a plausible dominance principle. So, while expectationism about utility may be implausible, once we move to an essentially comparative model, expectationism about utility-difference is more compelling.[201]

SSB is compatible with the Do Nothing verdict in the Easy Lottery. What of the other verdict, that you should uphold State-wise Dominance and so Pay to Permute? We can here draw on Fishburn's (1989b, 1990) *Skew-Symmetric Additive* (SSA) model. Again taking $\phi\colon O \times O \to \mathbb{R}$ to be unique up to positive scalar multiplication, define (assuming a countable state-space) the instrumental utility-difference Φ^* between f and g:

$$\Phi^*(f,g) = \sum_{s \in S} C(s) \cdot \phi(f(s), g(s))$$

This method of comparing acts begins with state-by-state comparisons and tells you to weight the utility-difference between acts in each state by the probability of that state. Trivially this rule satisfies State-wise Dominance, so it tells you to strictly prefer Paying to Permute in the Easy Lottery.[202]

Again, you might adopt a pluralist approach and think that both Φ (i.e., the SSB model) and Φ^* (i.e., the SSA model) represent legitimate ways of comparing acts. Or you might think that one will ultimately appear more attractive at the end of a process of reflective equilibrium. What

Fishburn (1982, Lemma 3) proves that for each act f, there exists a $\Phi_f(\cdot)$ such that $\Phi_f(g) \geq 0$ if and only if $g \succcurlyeq f$ and $\Phi(h) \leq 0$ if and only if $f \succcurlyeq h$; moreover Φ_f is expectational, that is $\Phi(\alpha_1 a_1 + \cdots + \alpha_n a_n) = \sum_i \alpha_i \Phi_f(\alpha_i)$. So, even without Symmetry we can for each act define a comparative expected utility function relative to that act (without Symmetry, however, the comparative utilities $\Phi_f(g)$ and $\Phi_g(f)$ need not be of equal magnitude).

[201] Note that the SSB model is compatible with Transitivity. Indeed, the SSB axioms plus Transitivity are equivalent to the axioms of Weighted Linear Utility Theory, which I discussed in Chapter 1. This highlights that an expected utility-difference model is compatible with risk-sensitive preferences and so permits a far more reasonable range of preferences than does the EU model.

[202] SSA is axiomatized in the Savage framework with remarkably few deviations from Savage's axioms. The STP is replaced with an explicit State-wise Dominance axiom, Transitivity is dropped, and the Archimedean axiom is tweaked. Fishburn (1989) provides the initial axiomatisation, which is extended to the case of a finite set of states in Fishburn (1990).

matters here is that both responses to the Easy Lottery can be vindicated by some formal decision theory; conversely, our judgements about the Easy Lottery provide tools with which to normatively assess novel formal models.

5.6 Conclusion

I have argued that Money Pumps do not undermine the case for rational cyclic preferences. We can rationalise cyclic preferences if we adopt a contrastive account of utility—we cannot make judgements about 'the' utility of an outcome, only the extent to which one outcome is preferable to another. From this perspective, the Money Pump highlights the costs of context-sensitivity in the dynamic setting: you must make distinct choices across different choice contexts, so you cannot pursue a unified strategy to arrive at a cost-free outcome. This cyclic preference tax can, however, be rationally borne.

I have also argued that risky cases like the Easy Lottery pose a challenge for defenders of intransitivity. There are several ways we might respond to these challenges, each with different normative upshots. While I here leave open various responses to the Easy Lottery, it is good to know that each response can be rationalised by some formal model (and in turn, thinking about the Easy Lottery highlights the normative underpinnings of those formal models). The SSB and SSA models provide distinct ways of formalising utility-difference or contrastive utility, and while there are significant differences in the foundations of those models, they provide an important first step towards a general decision theory without transitivity.

Chapter 6

Decision Without Transitivity

6.0 Preference Cycles and Dominated Choices

For the purposes of this chapter, I assume that (i) cyclic preferences are rational, and (ii) Skew-Symmetric Bilinear Utility Theory (SSB) is the correct rule for comparing risky acts. I have no argument for these assumptions beyond the discussion in the previous chapter. Nonetheless, they strike me as defensible working assumptions, and they provide a platform from which to tackle an important question: given cyclic preferences, what dominance constraints are there on choice?

Recall from the last chapter that SSB gives us tools with which to compare acts. Even if we reject Transitivity, we can compare acts as means to ends based on how likely those acts make outcomes. Moreover, we can assess not only which acts are preferable, but the strength of those preferences.

Now, it is uncontroversial that preferences constrain rational *pairwise choices*. This might be because preferences are choice dispositions (e.g., Savage 1972), because preferences are a kind of judgement that manifest in choice dispositions (e.g., Bradley 2017), or because preferences supervene on what you have most reason to choose (e.g., Thoma 2021). What matters here is that preferences do constrain pairwise choices—if they did not, then they would be the wrong focus for decision theory. Recalling that a *choice function* $\boldsymbol{c}(\cdot)$ maps each option set to a subset of permissible options, we know that:

Binary Choice: If $f \succ g$, then $f \in \boldsymbol{c}(\{f, g\})$ and $g \notin \boldsymbol{c}(\{f, g\})$.

So far, so good. Binary Choice does not assume Transitivity, so anyone who evaluates options in line with SSB can select a choiceworthy option in any binary decision.

But Binary Choice falls short as a general decision rule. The *general choice problem* is how preferences, which constrain pairwise choices, guide choice in situations with more than two available options.[203] Defenders of Transitivity have an easy solution to the general choice problem—they simply endorse (for finite option sets) Davidson et al.'s:

[203] I here make an assumption: preferences are primitive and so determine choice functions. Others might take choice functions as primitive and deny that preferences systematically constrain choices in non-binary sets. That I

Non-Dominated Choice: $f \in \boldsymbol{c}(A)$ if and only if there is no $g \in A$ such that $g \succ f$.

But defenders of cyclic preferences cannot straightforwardly endorse this constraint as the defender of Transitivity does. So, we must say more about how preferences constrain non-binary choices, or else our decision theory provides no action-guiding advice outside of a narrow range of situations.

I first consider two approaches that have been explicitly defended in the literature:

(i) M*aximal Choice*: expand option sets to include mixed-acts and retain the Non-Dominated Choice constraint (defended by Fishburn 1984a). I argue that this approach is *overly demanding.*

(ii) *Constrained Picking*: an option is impermissible only if it violates some minimal qualitative constraint (suggested by Herlitz 2020). I argue that this approach is *overly permissive.*

In response, I motivate and defend:

(iii) *Proxy Choice*: an option is permissible if and only if it maximises some quantity appropriately related to preference. I argue that this is the correct approach.

Finally, I consider a fourth response:

(iv) *Go Transitive*: adopt transitive preferences. I argue that while not required, this might be instrumentally beneficial for some agents, and the SSB framework provides novel tools for guiding this process.

While *Proxy Choice* solves the general choice problem, each other response tells us something important about the nature of choice without transitivity.

6.1 Maximality and Mixing[204]

So far, I have granted Gustafsson's (2013) claim that agents with cyclic preferences must violate the Non-Dominated Choice constraint. In a similar vein, Nebel (2018, p. 875) suggests that if 'for any outcome we choose or prefer, there may be some better alternative that we ought to have chosen or preferred instead', then this leads to 'skepticism about practical reasoning'. So,

opt for a preference-first, rather than a choice function-first, approach is a working assumption, albeit a substantive one.

[204] From here my focus is on choice, so I discuss preferences over acts, not outcomes. Any talk of outcomes should be interpreted as talk of the corresponding degenerate acts that yield those outcomes with certainty.

perhaps even if reasonable agents cannot be talked out of cyclic preferences, we cannot do decision theory with cyclic preferences.

Fishburn (1984a), however, argues that cyclicity is compatible with the Non-Dominated Choice constraint. Preference cycles might mean that each *pure* option is strictly dispreferred to an available alternative—if $a \succ b \succ c \succ a$, then deterministically selecting a single option from $\{a, b, c\}$ violates the Non-Dominated Choice constraint. But we often have access to more than pure options. Rather than deterministically selecting a single option, we can perform *mixed-acts*, which induce non-trivial probability distributions over those options (more on what this amounts to later). When confronted with a choice from set $\{a, b, c\}$, perhaps you should toss a coin or pick something at random.

You might think that if each option violates the Non-Dominated Choice constraint, the same goes for probability distributions over those options. But that is wrong, at least given setup so far. Fishburn (1984a, Section 6) notes that SSB meets the conditions required for the existence of Nash equilibria, meaning:

> **Mixed Existence:** For each option set A, there exist options $a_1, \dots, a_n \in A$ and probabilities $\alpha_1 \dots \alpha_n$ such that $a_1 + \cdots + \alpha_n = 1$ and $\alpha_1 a_1 + \cdots + \alpha_n a_n$ is weakly preferred to each option and probabilistic mixture between options in A.[205]

So, Fishburn (1984a, p. 81) suggests that we accept cyclic preferences and retain the Non-Dominated Choice constraint. This strategy yields some intuitive verdicts. Consider again Quinn's case:

> *Spectrum*: For any pain-level, you prefer gaining $\$10{,}000$ and increasing your pain-level by a fixed small increment, ϵ. Nonetheless, you prefer no pain and no gain over extreme pain with extreme gain.

Take three outcomes in this cycle, p_1 (the starting level of no pain), p_2 (some moderate amount of pain and moderate amount of wealth), and z (excruciating pain and extreme wealth). Assume $p_1 \succ z \succ p_2 \succ p_1$. We can now ask by how much each outcome is preferred to the next in the cycle. Though details will vary agent to agent, I suspect that my Φ has the following property:

[205] See Fishburn (1988, Section 3.11) for helpful discussion. Note that once we establish that a mixed-act is weakly preferred to each pure option, Mixture Dominance guarantees that it is weakly preferred to every probabilistic combination of those pure options. This indicates that Mixed Existence may not hold for non-transitive generalisations of Mixture Dominance-violating theories like Buchak's REU.

the utility-difference between p_1 (involving no pain) and z (involving extreme pain) is large compared to the utility-difference between outcomes closer together in the spectrum.[206] That is, something like the following holds:

$$\Phi(p_2, p_1) = \Phi(z, p_2) = 10 \text{ and } \Phi(p_1, z) = 100$$[207]

Given these utility-differences, it is easy to check that SSB ranks the mixed-act $\frac{1}{12}p_1 + \frac{1}{12}z + \frac{10}{12}p_2$ as weakly preferred to each pure option (and hence any probabilistic combination of them). So, what should you do when making a trade-off between pain and wealth? Since no single outcome is preferable to every other, you should not settle on a single outcome nor rule any such outcome out. Rather, you should randomise—in this case in a way that is highly biased towards the middling outcome. This makes intuitive sense: p_2 is not strongly dispreferred to anything, and it is preferred to p_1.

One objection: z, which involves extreme pain, is horrendous, so you should give it *no* weight! Since z is far worse than some available alternative, p_1, you might think it foolish to randomise and give yourself a chance of z when you could have p_1 outright. Note that in an analogous case from population ethics, Parfit's (1984) Spectrum Argument, which involves trade-offs in average well-being and population size, Parfit calls it *repugnant* to say that an enormous population of lives barely worth living is better than a small population consisting of excellent lives. Similarly in Quinn's case, you might think it repugnant to choose z when p_1 is available. And perhaps we should always avoid repugnant choices.[208]

In response, note that it is not z itself that is supposed to be repugnant. Rather, what is repugnant is the judgement that the terminal outcome in the spectrum is *better than* (or preferable to) the initial outcome. And the mixed-act strategy does not require us to say that z is better than

[206] MacIntosh (2010, p. 75) entertains a similar intuition.

[207] Such utility-differences might reflect a kind of primitive attitude on the agent's part that we derive by looking at their preferences over risky acts. Or they might arise from adopting some rule for comparing pain-wealth levels. For example, say that there is some threshold such that (i) increases in pain below that threshold are negligible, so only wealth-increases matter below the threshold, but (ii) increases in pain above that threshold are non-negligible, so render wealth-increases irrelevant:

> **Threshold Priority:** Say $x > y$. If the difference between x and y in terms of pain is below threshold t, then $\phi(x, y)$ is an increasing function of the wealth-difference between x and y, bounded by some real number M. If the difference between x and y in terms of pain is at least as great as t, then $\phi(x, y)$ is an increasing function of the pain-difference between x and y.

Note that when wealth matters, ϕ is bounded by M. So, if we let the magnitude of certain pain-increases be greater than M, then those pain-increases are of greater magnitude than *any* wealth-increase. Threshold Priority, which is a special case of a lexicographic semi-order (see Tversky 1969), is probably overly simple, but it shows the kind of essentially comparative structure that rationalises cyclicity. Thanks to Kirsten Mann for discussion here.

[208] Thanks to members of the ANU Decision Theory reading group for pressing this worry.

(or preferable to) p_1. Indeed, it allows us to say the opposite, that it is *much worse*. Nonetheless, z has something going for it that p_1 does not: it is an improvement on p_2. In this way, the mixed-act represents the optimal *ex ante* balance of each outcome's virtues and vices—the probability you assign z (which is dispreferred to p_1 by a lot) is precisely offset by the high probability of receiving p_2 (which is preferred to p_1). So, giving some weight to z is the best compromise between the competing demands of your cyclic preferences.

(As an aside, the Mixed Existence may make us question whether cyclicity really is a good explanation for the kinds of intuitions people report in spectrum cases. So far, I have been assuming that spectrum cases provide *prima facie* support for rational intransitivity. But we need to assess our intuitions about preferences holistically, and this means considering preferences over riskless acts in light of considered intuitions about preferences over risky acts. Some I have talked to steadfastly maintain that they would avoid *any* chance of excruciating pain in Quinn's case. If so, then this puts pressure on cyclicity as a rationalisation of their intuitions—if they really did have cyclic preferences, they would accept a non-zero probability of z. Personally, I find assigning z non-zero probability plausible, so I am happy to accept cyclicity as a rationalisation of my intuitions.)

Another objection: perhaps it is puzzling that whatever mixed-act you adopt, the outcome you receive at the end of the day is dispreferred to an available alternative. You may randomise, but *ex post* you will still think 'Blast! I could have got a preferable option outright'. So, while mixed-acts let us satisfy the Non-Dominated Choice constraint *ex ante*, they do not let us avoid dispreferred outcomes *ex post*.[209]

In response, the goal of decision theory is to guide, evaluate, or explain the behaviour of agents. This means that rational constraints always govern choices, not the outcome of those choices. Of course, the defender of cyclic preferences accepts that in some situations there is no 'best' or 'most preferable' outcome. This means that *ex post* regret might be inevitable. But the goal of decision theory is not to avoid *ex post* regret. Rather, since we are interested in guiding, evaluating, or explaining choices, what matters is that SSB recommends something rationalizable by your *ex ante* lights. And it does so by recommending mixed-acts.

So, mixed-acts can represent optimal trade-offs between the competing demands of cyclic preferences, in which case they are the best means to your ends. The reason I do not accept mixed-acts as a solution to the general choice problem is that rationality cannot demand that

[209] Thanks to Alan Hájek for pressing this worry.

mixed-acts are available as options.[210] For descriptive and modelling purposes, it might be useful to imagine option sets as containing all probability mixtures between pure options. But this is surely an idealisation when imported into a normative theory of individual rationality. Taking mixed-acts literally requires you to have access to some method for randomisation, which might be unavailable, costly, or time consuming. And rationality cannot require that you have access to chance devices that nature does not provide.[211]

Say that you do have a low-cost chance device. Even then there is the complication of saying what it means to act on the basis of that device. Say that you roll a die and it tells you to take p_2. Is this a mixed-act? Once the die has indicated p_2, you must now commit to p_2, say by forming an intention to bring it about. But that is not a mixed-act—you strictly disprefer choosing p_2 (even as the result of a die) to z. Once the die indicates an option, this is a non-chancy thing that you must do, moreover one that the Non-Dominated Choice constraint prohibits. So, even if you have access to a chance device, this does not entail the ability to act with non-trivial probabilities. (This is not to say that chance devices *never* enable us to act with non-trivial probabilities—imagine an app with a random number generator that performs tasks on your behalf. The claim here is just that rationality cannot demand that such devices are available.)[212]

[210] Meacham (2010, p. 66) similarly criticises decision theories that presuppose mixed-acts. Fishburn (1984, p. 81) claims that 'since convex combinations (probability mixtures) of feasible prospects are also feasible from an *ex ante* choice view-point, we consider them as part of the choice situation'. He does not, however, further explain this assumption. In what follows I argue that feasibility of prospects does not entail feasibility of probability mixtures of those prospects. Thanks to Christopher Bottomley, Wlodek Rabinowicz, and James Willoughby for helpful discussion of mixed-acts.

[211] Hájek (2014b) provides an ingenious method of generating any chance you like. But note that running the method he proposes takes time, which might be costly.

[212] Here is one possible response to the worry just raised. Recall that Machina (1989, p. 1647) thinks that agents who deviate from EU treat past risks as 'gone in the sense of having been *consumed* (or "borne"), rather than gone in the sense of *irrelevant*.' If so, then by randomising over acts, you may treat the randomly selected pure option as part of a probabilistic strategy. For example, if a die induces probability distribution $\frac{1}{12}p_1 + \frac{10}{12}p_2 + \frac{1}{12}z$ and subsequently recommends p_2, the fact that the die *could* have recommended p_1 or z means that, after the resolution of uncertainty, you evaluate p_2 in light of risks previously borne. So, you may treat p_2 as choiceworthy on the basis that you received it with probability $\frac{10}{12}$ and could have received p_1 or z with probabilities $\frac{1}{12}$. After the resolution of uncertainty, you still treat the pure option as part of a mixed-act whose risks were consumed in the past. This Machina-inspired solution does not guarantee that chance devices are available cost-free, and we might question whether rational agents are *always* motivated by the risks they bore in the past (if I can now choose between a million for sure and a 10% shot at billions, I suspect that I prefer the million *even if* I only face this choice after bearing risks in the past—give me the money, past risks be damned!). So, the Machina-inspired solution may not be available to all agents all the time. Nonetheless, if and when the strategy is viable, it greatly increases the scope of the mixed-act solution.

Similar problems arise if you think of mixed-acts as requiring not *external* chance devices but some internal stochastic process. That is, even if you think of mixed-acts as acting in line with internal 'choice probabilities', the problem does not go away.

Firstly, note that the probabilities involved in the mixed-act strategy might be very precise. While I might be able to choose red wine over white with probability $.5$, I am not sure what it means to choose red with probability $.7373$. (And if I suffer from 'trembling hands', there might be natural limits to the precision with which I can set probabilities of actions.) And even then, once I find myself reaching for red wine, I still must grasp the glass and drink.

Fundamentally, we ought to be able to give some account of what choice probabilities are. But it is often most natural to interpret choice probabilities as credences *that* you will choose something in a non-chancy way, not choosing something in a genuinely chancy way. For example, we might have non-trivial credence's in which act we will perform, but this is not the same as acting with non-trivial probabilities. Credences in one's acts feature prominently in, say, Arntzenius' (2008) and Joyce's (2012) variants on Causal Decision Theory. Take Joyce (2012, p. 134), who when discussing a case structurally similar to the Psychopath Button, says that given non-trivial credences in your acts, 'you can rationally choose to shoot [a pure option], to refrain from shooting [a pure option], or to perform any "mixed-act"'. Crucially, having non-trivial credences here does *not* thereby mean you perform a mixed-act—Joyce allows that you select a pure option on the basis of non-trivial credences. So again, cyclic preferences may require you to violate the Non-Dominated Choice constraint.[213]

But say that I can realise certain internal stochastic processes that make me act like a chance device. Then the same issues that problematised external chance devices re-arise. For example, say that I have an internal 'whim engine'. I am choosing between red and white wine on a whim, and I find my whim pointing to red, which it does exactly half the time. Am I really choosing red

[213] Arntzenius (2008, p. 292) analyses a 'mixed decision' as '[having] certain credences in one's acts at the end of rational deliberation'. He is careful to distinguish this from what I have been calling a mixed-act: 'mixed decisions are not decisions to *perform* certain acts with certain probabilities' (p. 292). Again, being able to perform a mixed decision in Arntzenius' sense does not allow the agent with cyclic preferences to satisfy the Non-Dominated Choice constraint. Even if an agent can adopt non-trivial credences in what they will choose at the end of some process of rational deliberation, they must still choose a non-chancy act at the end of that process, thereby violating the Non-Dominated Choice constraint.
I do note that Arntzenius (2008, 292) adopts view of decision theory not as recommending acts but as a theory of 'what credences one ought to have in one's actions'. If so, then we might say to the agent with cyclic preferences that they should simply adopt appropriate act-credences—so long as they have appropriate credences, there are no decision-theoretic constraints on what they actually end up *choosing*. I am not persuaded to adopt Arntzenius' proposal—it strikes me as shifting the goalposts to move from decision theory as a theory of *decisions* to a theory of *credences*. Nonetheless, someone who adopts Arntzenius' may be able to provide an account of deliberational dynamics that allows the agent with cyclic preferences to have appropriate 'credences in one's acts at the end of rational deliberation'.

with $.5$ probability? No—the problem with external chance devices now re-arises. Some stochastic process has taken place, the result of which is that I must now make an ordinary choice to have red wine. So, even if the whim engine assigns *ex ante* optimal probabilities over $\{p_1, p_2, z\}$, cyclic preferences mean that what I actually choose violates the Non-Dominated Choice constraint. Now, the 'whim engine' illustration is fanciful, but it does capture what goes on in many models of stochastic choice. For example, Regenwetter et al. (2010, 2011) interpret stochastic choice as involving your having a probability distribution over various mental states (which are complete preference orderings, or deterministic choice strategies). This means that though chance enters at the level of which mental state guides your choices, when you choose in line with a mental state that choice is not itself chancy.

So, I find myself puzzled about what mixed-acts are supposed to be. Perhaps in some cases we can literally behave like a chance device (just reach out and see whether you grab red or white wine!). And in others we might be able to outsource our decisions to chance devices. But in many cases, the thing we ultimately must do is select a non-chancy option, and if we uphold Davidson et al.'s Non-Dominated Choice constraint, this leaves us in a bind. So, I maintain that cyclic preferences require that we reject the Non-Dominated Choice constraint.

Nonetheless, the Mixed Existence result is still instructive. Situations in which we can feasibly defer to a chance device are likely situations where we have time, and situations where it is cost-efficient to do so are likely when randomising is cheap or the stakes are high. For example, imagine a group of friends ordering pizza. To overcome Condorcet-style voting worries, they download an app that directly places orders at local restaurants with non-trivial probabilities. They input their preferences into the app, knowing that it will order pizza with the appropriate probabilities. This really does seem like a chancy act that achieves an optimal *ex ante* balance between the friends' conflicting preferences. Or consider a policymaker who accepts a cyclic solution to the Repugnant Conclusion. They might well be able to create a weighted lottery and commit (e.g., through a legally binding procedure) to acting on the basis of the outcome of that lottery. So, provided we uphold the *sufficiency* half of the Non-Dominated Choice constraint—that it is always permissible to choose an option that is preferred to every available alternative—the Mixed Existence result is powerful. At least some preference cycles can be resolved by resorting to mixed-acts. And while this does not remove the sting from the general choice problem, it does drain some of the venom.

6.2 Constrained Picking

A second response to the general choice problem is to reject the Non-Dominated Choice constraint and replace it with some more minimal constraint. This typically involves specifying some pattern(s) of preference relations that render options impermissible, then saying that any option not thus rendered impermissible is permissible. One such constraint that I have already discussed is:

> **Non-Absolutely Dominated Choice:** $f \notin \boldsymbol{c}(A)$ if there is some $g \in A$ such that $g \gg f$.

Or we might insist on a still more permissive constraint:

> **Non-Fully Dominated Choice:** $f \notin \boldsymbol{c}(A)$ if for all $g \in A$ distinct from f, $g \succ f$.

Or we could go for a more restrictive constraint. Following Ahmed (2017, p. 998) we might say that in options set A (i) f is *weakly superpreferred* to g if for any $h \in A$, if $g \succ h$ then $f \succ h$, and (ii) f is *strictly superpreferred* to g if f is weakly superpreferred to g and g is not weakly superpreferred to f. We might then suggest:

> **Non-Super Dominated Choice:** $f \notin \boldsymbol{c}(A)$ if there is some $g \in A$ such that g is strictly superpreferred to f in A.

Various other 'minimum bars' for choice have been proposed in the literature (e.g., Herlitz 2020 considers ruling out options not contained in the Maximal Schwarz Set). The unifying theme behind this family of approaches is (i) qualitative information about preference relations rules some options out, and (ii) any option not thus ruled out is permissible. Since the Non-Absolutely and Non-Fully Dominated choice constraints leave at least one choiceworthy option in each set, even when the agent has cyclic preferences, this approach never leaves us in a bind.

This approach is, however, objectionably permissive. Since we rule options out based only on *patterns* of preference relations, we cannot discriminate between options in a 3-cycle.[214] This leads to some strange verdicts:

> *Fred*: Fred is choosing between Applies, Bananas, and Cherries and has a preference cycle $a \succ b \succ c \succ a$. He is willing to pay a dollar to trade Apples for Bananas, Bananas for

[214] This criticism applies also to Brown's (2020) *Transitive Closure Maximisation* rule, for discussion (and compelling critique) of which, see Mann (forthcoming).

> Cherries, and Cherries for Apples. When all three options are on the table, he is happy to have any of them.
>
> *Patricia*: Patricia is choosing between three pain-wealth levels p_1, p_2, and z and has a preference cycle $p_1 \succ p_2 \succ z \succ p_1$. Patricia would pay a large sum of money to move from z to p_1 (far larger than the sum she would pay to move from p_1 to p_2 or p_2 to z).

Constrained picking says that Fred and Patricia are in identical positions. The sets $\{a, b, c\}$ and $\{p_1, p_2, z\}$ contain the same patterns of preference relations—each option beats and is beaten by exactly one other. But surely our theory can offer different advice to Fred and Patricia![215] Fred faces a symmetric cycle, so there is no relevant difference between outcomes that he can use to disqualify one but not another. But Patricia faces a highly asymmetry cycle, and there is a strong intuition that, if she must pick an option outright, she should rule out z.[216] But the constrained picking approach cannot accommodate this verdict.

It is worth quoting here Herlitz who, in the context of population ethics, suggests a version of constrained picking. He claims:

> 'Accepting a cyclical ranking of different alternatives and pairing this with a revision of one's conception of rational choice in the standard Spectrum Arguments will, after all, deem many options permissible (including world Z in the Repugnant Conclusion and curing headaches in the healthcare example), and this seems wrong. [Note: choosing world Z in the Repugnant Conclusion corresponds to choosing z in Quinn's Self-Torturer.] The view will entail that a world with a huge population of people leading lives barely worth living is permissible even if there is an alternative world with a large population of very happy people, and the view will entail that it is permissible to provide minor health benefits to a large group of people even if one could instead save some other person's life. This seems intuitively wrong.
>
> To this objection, one can first point out that there is no view in the literature that renders judgments that are in line with widely held intuitions in all cases.' (Herlitz, 2020, p. 186)

[215] Brown (2020, p. 51) argues that Fred and Patricia are in identical positions because each option they face is 'on a par' (each beats and is beaten by one). But to say that the options are on a par simply assumes that preference patterns, not strength of preference, are all that matter. The intuition against choosing z outright strongly indicates that each option is *not* on a par for Patricia (or at least that she can reasonably look for a decision rule that does not say they are on a par). MacAskill (2016, p. 993), in the context of normative uncertainty, makes a similar claim that 'it is difficult to see what a choice-worthiness cycle could contribute other than that all options within the cycle contribute equally to the appropriateness ranking'; but again, this assumes that patterns of relations, rather than strengths of relations, are all that matters.

[216] See Mann (forthcoming) for discussion and defence of a parallel claim in moral cases. MacIntosh (2010, pp. 75) entertains similar intuitions.

So, Herlitz accepts the intuition: some options are distinctively bad in a cycle. His response is that no view in the literature accommodates all widely held intuitions, so this is the best we can do. But we now have richer evaluative resources with which to accommodate widely held intuitions. SSB *can* vindicate the intuition that there is something distinctively wrong with choosing z outright. And it can do so without giving up on cyclicity, so can dodge the repugnant conclusion that terminal outcomes in spectra are preferable to starting ones. In the next section I formulate rules that take account of utility-differences and rule out z. So, if no view in the literature accommodates all widely held intuitions, this is because the literature has not paid sufficient attention to quantitative models like SSB.

Tellingly, Herlitz goes on to say that the overly permissive nature of constrained picking speaks against a purely axiological approach to choice (i.e., one that takes betterness, or in this case preference, relations as the only constraint on choice). He therefore suggests (p. 187) that we may need to *complement* decision theory and go beyond facts about preference/betterness to explain why people rationally reject z in spectrum scenarios. Since facts about utility-difference are about preference/betterness, the next section demonstrates that we need not go beyond the standard picture of decision theory, as Herlitz suggests, to arrive at plausible results.

6.3 Choice by Proxy

We need some quantity appropriately related to preference that guides choices. I argue that *negative utility-difference* is an appropriate quantity and that *minimise negative utility-difference* is a reasonable decision rule. I will not argue that this is the only plausible rule—I leave open that there are a range of permissible ways to reason when confronted with cycles. Nonetheless, the rule I discuss serves as proof of concept that preferences can guide choice in a way that vindicates a core intuition: even if I have cyclic preferences, I should avoid z when selecting an option outright.

Consider again $p_1 \prec p_2 \prec z \prec p_1$. I always assume that $\Phi(p_1, z)$ is the greatest utility-difference between options. When we judge z to be impermissible, the simplest explanation is that we eliminate the option with the greatest *negative utility-difference*, which motivates:

Least Bad: $f \in \boldsymbol{c}(A)$ if and only if for all $g \in A$: $\min_{x \in A} \Phi(f, x) \geq \min_{x \in A} \Phi(g, x)$.

We can think of this as saying that, while you may not be able to fully satisfy your preferences, you can minimise the extent to which you violate them. And since each option's negative utility-difference is a real number, $\min_{x \in A} \Phi(a, x)$ acts as a kind of proxy value for a in A. For example, in our canonical case $\Phi(z, p_1) = -100$, so -100 is the proxy value for z in the set $\{p_1, p_2, z\}$. Since p_1 and p_2 have proxy values -10, Least Bad rules out z. This is a sensible form of practical reasoning given cycles (contra Nebel 2018, p. 875) that does not require anything beyond facts about preference/betternesss (contra Herlitz 2020, p. 187).

I will not here make the strong claim that Least Bad is the only plausible rule that accommodates cyclicity. My goal in this chapter is proof of concept: we can uphold a range of reasonable intuitions and rationalise them by the lights of a satisfactory decision rule. At the very least, however, Least Bad satisfies two minimal but non-trivial constraints.[217]

(i) Least Bad satisfies the *sufficiency* half of the Non-Dominated Choice constraint.

That is, Least Bad satisfies:

Sufficiency of Non-Dominated Choices: If for all $g \in A$, $f \succcurlyeq g$, then $f \in \boldsymbol{c}(A)$.

I cannot think of a view in the literature that violates this constraint. And rightly so—you have no reason against choosing an option weakly preferred to everything else, so you are permitted to choose that option.[218]

Nonetheless, given cyclicity, the Sufficiency of Non-Dominated Choices is a non-trivial requirement. For example, consider an agent who chooses options that do *best* with respect to some other:

Most Good: $f \in \boldsymbol{c}(A)$ if and only if for all $g \in A$: $\max_{x \in A} \Phi(f, x) \geq \max_{x \in A} \Phi(g, x)$.

This violates the Sufficiency of Non-Dominated Choices in:

[217] Quiggin (1994) analyses rules with some similarities to Least Bad. I will not consider his approach here because it specifically adopts Loomes and Sugden's (1982) *Regret Theory* as a framework for intransitivity and assumes that outcomes are transitively ordered. As such, Quiggin's analysis solves various problems for agents who exhibit 'higher-order intransitivity'—transitive preferences over riskless acts with intransitivity arising when we bring risk into the picture. Merely higher-order intransitivity does not capture what is going on in riskless cases like Quinn's Self-Torturer, so I set it aside here.

[218] To see that Least Bad satisfies this constraint, note that if $f \succcurlyeq x$ for all $x \in O$, then $\min_{x \in O} \Phi(f, x) = 0$. And since for all g, $\Phi(g, g) = 0$, 0 is the best possible Least Bad score. Note that when mixed-acts are available, this means that Least Bad always permits the optimal mixed-act guaranteed by the Mixed Existence result. So, Least Bad says that optimal mixed-acts, when available, are choiceworthy because they are what you have least reason *not* to do.

> *Bad Cycle*: You have preferences $a \succ b \succ c \succ a$, and there is some d such you strictly prefer d to each of $\{a, b, c\}$. You prefer each element in the cycle strongly to the next: $\Phi(a, b) = \Phi(b, c) = \Phi(c, a) = 10$. Your preference for d over each is not quite as strong: $\Phi(d, a) = \Phi(d, b) = \Phi(d, c) = 5$.

For example, maybe you have ambivalent tastes about music (you strongly prefer each of a, b, c to another), but dislike music in general. So, you would rather go for a pleasant drive, d, than listen to any music. But Most Good says that your ambivalence about music means you should listen to music, even though you really want to go for a drive! That is extremely odd. Indeed, you prefer d to a, so why choose a? By choosing a, you choose something that you take to be worse for you than d. And there are no drawbacks to moving to d instead of a: d beats everything that a beats. So, you have no all-things-considered reason not to choose d over a. The same goes for b and c—d beats both, and you have no reason not to take d. So, Most Good violates the Sufficiency of Non-Dominated Choices, objectionably ruling out options that you have no reason not to take.

Indeed, we can go further and diagnose the fallacy that leads Most Good to make these odd recommendations. It wrongly treats the fact that x beats y as a reason for taking x when y is available. This commits what Muñoz calls the 'Second Cookie Fallacy', which he illustrates using the following case:[219]

> *Two Cookies*: I steal your cookie. You ask me to justify myself—what reason did I have for stealing a cookie? I provide the following reason: I could have stolen *two* cookies! So, though it might not have been decisive, there was a fact that counted in favour of stealing a cookie, that it was better than some available alternative.

This is surely fallacious—if you disagree, can I ask whether you have any cookies? There are stronger reasons against stealing two cookies than taking one (and stronger reasons against stealing no cookies than taking any). But this fact does not itself speak in favour of stealing a cookie. On the other hand, not stealing a cookie is the right thing to do simply because no option is better. This suggests an asymmetry in our ordinary practical reasoning: it counts *against* stealing a cookie that it is worse than an alternative, but it does not likewise count *for* stealing a cookie that it is better than an alternative. Rules like Most Good evaluate options by what they

[219] Case used with permission. Thanks to Daniel Muñoz for helpful discussion here.

beat, which as the Two Cookies case shows, fails to respect a plausible feature of our ordinary practical reasoning.

Least Bad, by contrast, rationalises a common-sense way of thinking about Two Cookies: I could do better than steal a cookie, so I have a reason not to do so; I could do *far* better than stealing two cookies, so I have even more reason not to do that; and there is nothing better than not stealing, so that is the only thing to do. You get closer to the choiceworthiness medal insofar as you are beaten less by other runners. You do not get closer to the medal because someone worse than you joined the race.

So, Least Bad satisfies one minimal but non-trivial dominance principle, the Sufficiency of Non-Dominated Choices, which reflects the way that it encodes a style of common-sense form of practical reasoning.

(ii) Least Bad satisfies the Irrelevance of Indiscernibles.

Another minimal constraint on a reasonable decision rule is that it satisfies:

> **Irrelevance of Indiscernibles:** Say that for $f \in A$ there exists $g \notin A$ such that for all $h \in A$, $\Phi(f, h) = \Phi(g, h)$. Then $f \in \boldsymbol{c}(A)$ if and only if $f \in \boldsymbol{c}(A \cup g)$.

This says that if g is a 'copy' of f—identical with respect to every comparison you might make with other options—then its presence (or absence) in an option set does not affect the permissibility of the other options. Again, this captures a compelling feature of our ordinary practical reasoning. Merely finding out that one option can be realised in two identical ways should not change what you take to be appropriate means to your ends.[220]

Again, the Irrelevance of Indiscernibles is non-trivial given cyclic preferences. For example, consider a rule that evaluates options based on the *total amount* by which they are beaten by others:

> **Aggregate Badness:** $f \in \boldsymbol{c}(A)$ if and only if for all $g \in A$:

$$\sum_{x \in A : x \succ f} \Phi(a, x) \leq \sum_{x \in A : x \succ g} \Phi(b, x).$$

[220] To see that Least Bad satisfies the Irrelevance of Indiscernibles, note that if g is a copy of $f \in A$, then for all $y \in A$, $\min_{x \in A} \Phi(y, x) = \min_{x \in A \cup g} \Phi(y, x)$.

This violates the Irrelevance of Indiscernibles in cases like:[221]

> *One Haggis*: You are choosing between f (a fruitcake), g (a galette), and h_1 (a haggis). Your utility-differences are: $\Phi(f, g) = 10$, $\Phi(g, h_1) = 100$, and $\Phi(h_1, f) = 1$. You choose f (which is recommended by both Least Bad and Aggregate Badness).
>
> *Multiple Haggises*: The case is as above, except you have now learnt that there are a thousand available haggises, for each of which $\Phi(h_i, f) = 1$, $\Phi(g, h_i) = 100$, and $\Phi(h_i, h_j) = 0$.

We now get that the aggregate badness (AB) of f is:

$$AB(f) = \sum_i \phi(h_i, f) = -1{,}000$$

While:

$$AB(g) = -10$$

And for all h_i:

$$AB(h_i) = \Phi(h_i, g) = -100$$

So Aggregate Badness rules out the fruitcake when multiple identical haggises are available, though it permits the fruitcake when only one is available. This violates the Irrelevance of Indiscernibles. Again, this is odd. You have a weak reason to take each *individual* haggis over the fruitcake in Multiple Haggises, one that is far weaker than your reason not to take any haggis over the galette. Why treat these multiple, distinct weak reasons as if they provided a strong reason? My flat is slightly dirtier than the flat above me, so I'd pay $10 to move upstairs; on learning that there are two near-identical ways of leasing the upstairs flat, I have no stronger reason to move upstairs (I would pay no more to move, and I feel no more pressure to move). This suggests that our ordinary practical reasoning treats it is an *aggregative fallacy* to combine the reasons provided by multiple, distinct options into a single reason. Least Bad again respects this plausible feature of our practical reasoning and so upholds the Irrelevance of Indiscernibles.

[221] The following case puts pressure on qualitative analogues of Aggregate Badness. For example, MacAskill (2016) proposes a version of the Borda count in which each option f is evaluated by its choiceworthiness score: +1 for each option such that $f \succ g$, -1 for each option such that $f \prec g$, and +0 for each option such that $f \sim g$. Criticisms of Aggregate Badness apply to this Borda count rule.

So, Least Bad permits cyclic preferences and satisfies two extremely plausible but non-trivial constraints on choice—the Sufficiency of Non-Dominated Choices and the Irrelevance of Indiscernibles. It also vindicates the intuition that you ought not choose z in spectrum cases. You might want more from your decision rule, so you might look for further constraints that complement or revise Least Bad. At this stage, I am unsure whether we should expect to be able to pin down a unique rational decision rule for agents with cyclic preferences. We might rather think that, just as there are multiple permissible risk-attitudes that meet a core set of constraints, there are multiple permissible ways of reasoning about option sets that meet a core set of constraints. Settling that question is for future work. What matters here is that Least Bad is proof of concept that cyclicity does not force us to give up on practical reasoning, and it helps us get clear on which dominance principles constrain choice.

6.3.1 Fully-Dominated Options

Least Bad satisfies a number of plausible constraints. It also violates one that you might have thought was non-negotiable, which is:

Non-Fully Dominated Choice: If for all $g \in A$ distinct from f, $g \succ f$, then $x \notin \boldsymbol{c}(A)$.

Surely, you might think, if an act is *strictly* dispreferred to *every* available alternative, then that act is impermissible. A fully dominated option leaves you maximally dissatisfied—anything else you could have chosen would be preferable.

To see that Least Bad violates the Non-Fully Dominated Choice constraint, consider $e \succ f \succ g \succ e$, with h dispreferred to each. Let $\Phi(e,f) = \Phi(f,g) = \Phi(g,e) = 10$ and $\Phi(e,h) = \Phi(f,h) = \Phi(g,h) = 1$. Least Bad here says that you ought to choose the fully-dominated h from $\{e,f,g,h\}$.

There are two possible responses given the position I have adopted so far. The first is to *supplement* Least Bad with some qualitative constraint. For example, we could say that you choose the Least Bad option out of the subset of non-fully-dominated options. Herlitz (2020, p. 180) does accept the possibility that qualitative constraints might 'be amended with some view of how to make all-things-considered reasonable choices after the impermissible alternatives have been discarded'. If so, then we can view Least Bad as the necessary amendment:

Least Bad*: $f \in \boldsymbol{c}(A)$ if and only if it minimises negative-utility difference in A^*, where $A^* = \{x \in A: \exists y \in A, x \succcurlyeq y\}$.

We can further debate whether Non-Fully Dominated choice or some stronger constraint (e.g., Non-Absolutely Dominated choice) is the right tool with which to identify impermissible options. So long as that constraint entails Non-Fully Dominated Choices, Least Bad can be amended to satisfy our minimal dominance constraint.

The second response is to reject even the Non-Fully Dominated Choice constraint. I think this approach is defensible. I have already argued that it can be permissible to choose an option that is dispreferred to *an* alternative. So, it might be permissible to choose an option that is dispreferred to *all* alternatives.

Recall my justification for the Sufficiency of Non-Dominated Choices: it is always permissible to choose an option that is preferred to every alternative, because you have no reason *not* to choose such an option. There is no parallel justification, however, against choosing an option strictly dispreferred to every alternative. In the case just given, none of e, f, g represents something that you should obviously choose over h. For example, though e is strictly preferred to h, you have strong reasons against taking e (it is far worse than g). Similarly for f and g: though each is preferred to h, both are strongly dispreferred to something else. So, given the structure of the entire option set you face, none of $\{e, f, g\}$ stands out as the thing to do instead of h. Therefore, a rational agent may look at h and see it as an appropriate compromise given the structure of their option set.[222]

When we think about concrete cases, intuitions in favour of the Non-Fully Dominated Choice constraint become even weaker. Consider again our motivating case $p_1 \succ z \succ p_2 \succ p_1$ and the optimal mixed-act $m = \frac{1}{12}p_1 + \frac{1}{12}z + \frac{10}{12}p_2$. Now, perhaps m is not an option—mixed-acts cost some small fee. No matter, let m^* be m minus that small fee, $\$\epsilon$, and assume that for all outcomes x, y that $\phi(x, y - \$\epsilon) > \phi(x, y)$ (this just says that subtracting a small fee from the

[222] Note that if options are transitively ordered, there will be an option in each set that is weakly preferred to everything else, so whose Least Bad score is 0. A fully dominated option has a Least Bad score greater than 0, so Least Bad entails the Non-Fully Dominated Choice constraint in the special case of a transitively ordered option set.

second outcome always strengthens your preference for the first). It follows that each of p_1, p_2, z are preferred to m^*.[223] So, m^* is fully dominated in the option set $\{p_1, p_2, z, m^*\}$.

Yet it is easy to rationalise choosing m^*. The optimal mixed-act m represent the best *ex ante* trade-off between each pure option's virtues and vices. The slightly sub-optimal mixed-act m^* still represents a very good *ex ante* trade-off between those virtues and vices. By taking m^* I avoid selecting something that is much worse than an available alternative—for each dispreferred outcome, I introduce some offsetting probability of a preferable option. So, from an *ex ante* perspective, m^* represents the best (well, least bad) I can do to balance the competing virtues and vices of each option.

Again, the fact that I will experience *ex post* regret need not deter me from taking m^*. I might end up with, say, $p_2 - \$\epsilon$ when I could have p_2 outright. But the goal of decision theory is not to avoid *ex post* regret. By taking p_2 outright, I would do something significantly worse than an available alternative, z. In contrast, by taking m^* I give myself some probability of z, along with balancing chances of p_1 and p_2. Cyclicity again puts us in a strange position where what seems best *ex ante* is *ex post* regrettable. But that is just the kind of strangeness I think we might learn to live with—indeed, as I argued in the previous chapter, by adopting the SSB framework we implicitly adopt a picture of rationality on which the *ex ante* perspective is action-guiding and so trumps the *ex post* perspective.

By permitting fully dominated options like m^*, the Mixed Existence result now becomes far more interesting. Mixed-acts may not always be available cost-free. But when the cost is low enough, Least Bad will recommend paying the required fee to randomise. So we can, while not fully satisfying our preferences, in a wide range of circumstances approximate an optimal resolution to the conflicting demands of cyclic preferences.

We therefore have two ways to think about the Non-Fully Dominated Choice constraint. We can either accept it as a side-constraint that supplements Least Bad, or we can reasonably reject it. I personally see no problem with rejecting it. We might resort to a kind of pluralism here (you might supplement Least Bad, while I may not). Or further discussion might push us towards a

[223] For example, $\Phi(p_1, m^*) = \frac{1}{12}\phi(p_1, p_1 - \$\epsilon) + \frac{1}{12}\phi(p_1, z - \$\epsilon) + \frac{10}{12}\phi(p_1, p_2 - \$\epsilon)$. Recall that $\Phi(p_1, m) = \frac{1}{12}\phi(p_1, p_1) + \frac{1}{12}\phi(p_1, z) + \frac{10}{12}\phi(p_1, p_2) = 0$. Since $\phi(p_1, p_1) < \phi(p_1, p_1 - \$\epsilon)$, $\phi(p_1, z) < \phi(p_1, z - \$\epsilon)$, and $\phi(p_1, p_2) < \phi(p_1, p_2 - \$\epsilon)$, we get that $\Phi(p_1, m^*) > \Phi(p_1, m) = 0$.

unique rule. Either way, Least Bad is a promising foundation for a solution to the general choice problem.

6.4 Going Transitive

A final approach to choice without transitivity is defended by Voorhoeve and Binmore (2006). They claim that agents with cyclic preferences should simply adopt (or reason their way to) a set of transitive preferences. Now, it might be puzzling that after defending rules that accommodate cyclic preferences, I consider this as a fourth option.

I do so because while intransitive preferences may be reasonable, they may also incur costs. Recall Rabinowicz's (2014) conditional recommendation: *if* you want to avoid the costs of disunified decision making, then you had better adopt non-exploitable preferences. We can therefore accept that some rational agents pay the cyclic preference tax—the structure of your tastes, goals, and desires is not subservient to tax avoidance—but that others might wish to avoid exploitation. If you are unsure about the frequency of opportunities for exploitation, or are convinced that many such opportunities will arise, then you might adopt transitive preferences as a precautionary measure. So, can we say anything about what kinds of agents in what kinds of situations will prefer to go transitive?

Hammond's (1988) famous consequentialist argument shows that, under conditions of dynamic separability, only EU-maximisers avoid exploitation.[224] So, we should expect non-EU-maximisers to recognise the potential for at least some inefficiency.

But when precisely will a non-EU-maximiser, in particular one with cyclic preferences, think it all-things-considered best to adopt non-exploitable preferences? I doubt that we can say anything terribly precise in answer to that question. Firstly, it involves a judgement about every possible decision situation the agent may face in the future, which I doubt is the kind of thing many of us can assign precise probabilities to. Secondly, it involves the question of how to evaluate bundles of goods (say, saving a dollar in a Money Pump one day *then* choosing Dylan over Bach the next day). And preferences over such combinations of goods are themselves a matter of taste that are not settled by our preferences over the individual goods. Finally, undergoing preference change might very well involve becoming a different sort of person. And, while this is not the place to

[224] Dynamic separability is the requirement that what is choiceworthy at some time supervenes on the structure of the decision situation from that time onwards (in particular, choiceworthiness now does not depend on facts about past decisions or risks). I question this assumption shortly.

delve into the extensive literature on decision theory and transformative experience, I am suspicious of norms that require you to undergo a transformative experience. For example, say that I have a preference cycle $\mathsf{Bach} \succ \mathsf{Dylan} \succ \mathsf{Beethoven} \succ \mathsf{Bach}$ because of my strong views about music; if I had to give up or moderate some of those views, I might think of myself as becoming a different kind of person with different values. And it is unclear that I consider myself better off all-things-considered by becoming a different kind of person. My preferring to save a dollar does not entail my preferring to save a dollar *by* adopting different preferences. So, I am suspicious of any sharp demarcation between agents who should and should not undergo preference changes. Whether to adopt non-exploitable preferences strikes me as a matter personal, subjective judgement.

One thing that we can say is that Least Bad by itself is highly context sensitive—which options are permissible is highly sensitive to the alternatives in the option set. Subtracting an impermissible option from a set may change what is choiceworthy, as can adding an impermissible option. That is just to say that Least Bad violates Sen's (1971) expansion and contraction consistency principles, which is not by itself a problem since accepting cyclicity essentially amounts to rejecting those principles (cf. Anand 1993, p. 339).[225] Nonetheless, we might hypothesise that as agents face more and varied choice situations, the opportunities for inefficiency become greater, which in general increases the pressure to adopt non-exploitable preferences. For example, consider:

> *Judgemental Jane*: Jane prefers Abba to Bach, Bach to Chopin, and Chopin to Abba. Jane recognises that in principle this leaves her open to exploitation. But Jane's tastes are deeply engrained. This means that she balks at violating any of her current preferences (far more than she balks at losing a dollar). And anyway, Jane rarely gets time to listen to music, so she has a relatively small number of decisions to keep track of and relatively few actual opportunities for exploitation.

Jane may risk paying taxes for a good cause, as I have argued. But contrast Jane with:

> *Procrastinating Pete*: Pete has the same preferences as Jane but finds himself falling prone to grievous procrastination. He is always trading records and ending up worse off.

[225] Caveat: if mixed-acts are available, then our choice function satisfies Sen's principles α and γ, though not β (see Fishburn 1984, p. 81). So, a moderate contraction and expansion consistency governs which acts you take to be appropriate means to your end(s). Of course, if those optimal means in the form of mixed-acts are not available as options, then the output of our choice function is much more context-sensitive.

> Unable to settle on a single musician to listen to, he finds himself constantly frustrated. Over time, Pete begins to wish that he could be like his transitive friends.

An adequate theory of rationality ought to say something to Pete as well as Jane. This is where we might turn to Voorhoeve and Binmore (2006), who claim that careful reflection will lead rational agents to arrive at transitive preferences. They claim:

> By a process of jockeying—making judgements under different presentations, checking their consistency, questioning inconsistent judgments and their grounds, discarding orderings resulting from undependable presentations and methods of evaluation in some cases, revising her judgments in others, again checking their consistency, etc.—[the intransitive agent] should ultimately arrive at an ordering that is consistent across different (non-misleading) presentations that respects transitivity. (Voorhoeve and Binmore 2006, p. 112)

Now, the discussion so far puts pressure on this general claim. The existence of a coherent SSB function indicates that Pete might carefully reflect on his preferences and still not arrive at a single, transitive preference ordering. Where Voorhoeve and Binmore think that a process of jockeying will compel rational agents to reject some of their preferences (or see that their true preferences were transitive all along), I maintain that rational agents might jockey and simply see that some utility-difference function Φ rationalises their essentially comparative judgements. Of course, Voorhoeve and Binmore might be right that some agents (appear to) have cyclic preferences because they employ unreliable heuristics for comparing options, which disappear when agents jockey in the manner described above. But this leaves open that other agents reflect carefully on their preferences, see that they do have cyclic preferences, and cannot identify any of those preferences as unreliable or worthy of rejection. So, it is entirely possible that someone like Pete jockeys in the manner described but finds himself stuck with exploitable preferences.

Here is where the SSB framework provides a powerful tool for Pete. Recall that Φ is bilinear, meaning that for every act f there is a one-place *linear utility* function $\Phi_f(\cdot) = \Phi(\cdot, f)$, which for each act assigns an expected utility relative to f.[226] Φ_f does not represent preferences, rather preferences relative to f:

$$\Phi_f(a) \geq \Phi_f(b) \text{ if and only if } \Phi(a, f) \geq \Phi(b, f)$$

[226] In fact, recall from Footnote 203 that Continuity and Mixture Dominance suffice—this result holds without the third SSB axiom, Symmetry.

Now, if Pete can pick some *reference act* r and evaluate each act based on how it compares with r, then he can behave like an Expected Utility maximiser relative to Φ_r. And, since EU-maximisers are dynamically consistent, Pete is invulnerable to exploitation.

How should Pete go about picking such a reference act? I do not think there is a single way he must go about this—I do not believe in hard and fast constraints on tastes, so I do not think there is some way that Pete *must* re-work his tastes. Instead, he will be guided by what seems right to him. Perhaps Pete thinks about a normatively salient point (e.g., if he is weighing lives, he might take r to be bringing about a neutral life). Or perhaps there is some psychologically salient point for Pete (e.g., in the case above, not listening to music at all). Or perhaps Pete must simply pick an option at some time and subsequently takes that to be his reference act. In this way, Procrastinating Pete becomes:

> *Proactive Pete*: I recognise that I am in an environment where intransitivity is costly. I therefore want to engage in tax avoidance, but I cannot shake my intransitivity simply by reflecting on my preferences. So, I will judge listening to a record to be good insofar as it is better than not listening to any record at all (or, in the case of Cher, to be bad insofar as it is worse than not listening to any record at all).

We might think of this as a kind of guided preference change on Pete's part. He reasons his way to EU-maximisation not by ignoring his preferences, but by taking certain features of his preferences (comparisons with 'No music') to form the right kind of basis for action. And because of the linear structure of Φ_r, this results in dynamic consistency.[227]

To reiterate, I do not think that rationality *requires* Pete to undergo this kind of process, nor do I think that every rational agent must be able to engage in this kind of process. I have assumed that Pete has some level of control over his preferences, can take certain features of his preferences as the right basis for action, or can move himself to compare options in a certain light. Plausibly, such an ability is one that some of us have some of the time. But when we do not, then following Bradley (2017, p. 286), we may have to do the best we can with the preferences we are endowed with. So, contra Voorhoeve and Binmore, I do not think that

[227] There is an overlap between the way I am using the term 'reasoning' here and the way that Broome (2013) does. We both focus on reasoning as a kind of conscious mental process that moves you from one set of preferences (or attitudes) to another. We also both think that reasoning involves following rules that 'seem right' to you (see Broome 2013, pp. 237-238); in my case, this is because choice of a reference act will ultimately be guided by what 'seems right' to you. We could therefore think of the reference act solution as providing a tool that real-world agents can use to guide reasoning. Of course, Broome is largely interested in reasoning from *inconsistent* attitudes to consistent ones. By contrast, I think that cyclic preferences are coherent so am thinking of reasoning as reworking a set of consistent preferences to achieve some goal (such as avoiding exploitation).

rational agents can simply reflect on their preferences and go transitive. But the discussion so far provides a powerful tool for guiding preference change when possible. Rather than simply telling agents to reject some of their preferences, reference-act dependent EU functions allow us to attend to specific features of our existing preferences and leverage those features to arrive at new ones (that satisfy the goal of invulnerability to Money Pumps).

6.5 Don't Look Back?

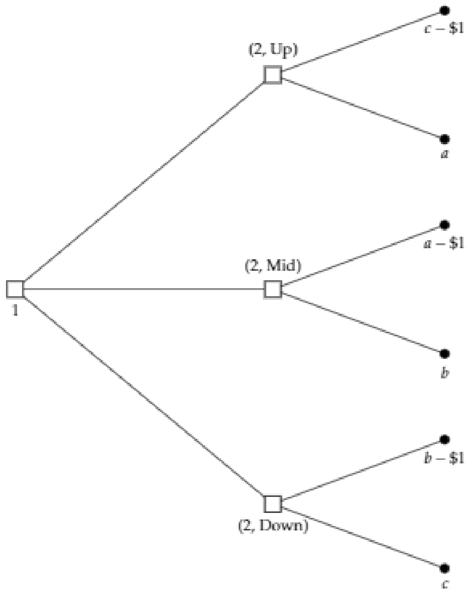
I finally want to address the elephant in the room: what kinds of choice strategies can you adopt to minimise (or avoid) the cyclic preference tax? Having viable tax-avoidance strategies minimises the risk of exploitation, so the instrumental reasons against exploitable preferences will vary with the viability of such strategies. There has been a lot of philosophical discussion around *forwards-looking* strategies—for example Sophisticated Choice (introduced in Hammond 1976) and Ahmed's Self-Regulating approach (2017). I here discuss the possibility of *backwards-looking* strategies—those that involve making choices based on what you had or could have had in the past.[228]

Take a preference cycle $a \succ b \succ c \succ a$. Say that you face a choice between a and b, that you choose a, and are then confronted with a choice between keeping a or swapping it for c. Should you assess this second choice relative to the set $\{a, c\}$ or, recognising that you just passed up b, the set $\{a, b, c\}$? Loomes and Sugden (1987, p. 286) suggest the latter.[229] And indeed, it strikes me as psychologically plausible that the *context* relative to which we evaluate options depends on facts about what we could have had.

[228] In this section, I deal exclusively with deterministic decision situations. Following Cantwell (2003, p. 385, notation adjusted) a deterministic decision situation is characterised by a finite set of *choice nodes*, N. For each $n \in N$, there is a finite set of *choices*, A_n, you can make at that node, where each $a_i \in A_n$ is a function mapping n to another node, $a_i: N \to N$. When A_n is empty, assume that n is an *outcome* (in the Savage sense). A sequence is an ordered tuple of choices $< a_1, \dots, a_n >$ such that a_1 is available at some node n_1, each a_{i+1} is available at the node $a_i(a_{i-1}(\dots(a_1(n_1)))$, and the final choice maps to an outcome. I assume that each decision tree has a starting node (i.e., every sequence begins with a choice available at the same node n_1). Informally: you can make various sequences of choices, each moving you from one choice node to another, the end result of which is an outcome.

[229] They back this up by claiming (p. 286) that after choosing between a and c, the agent will look back at the *entire* sequence of choices they made and experience regret or rejoicing relative to the set of options that were feasible for them at some time. *Ex post* regret and rejoicing enters because of Loomes and Sugden's psychologistic analysis of intransitivity as arising from 'basic utility' moderated by *ex post* feelings of regret. Such an analysis is markedly different to the one I have offered, which is in terms of essentially comparative *ex ante* evaluations. Nonetheless the basic point, that agents evaluate options relative to what they could feasibly have had at some time, is plausible even if we reject Loomes and Sugden's specific analysis.

Agents who choose in this way may avoid selecting dominated sequences of choices. Consider again Cantwell's Money Pump:



The set of possible final outcomes here is $\{a, a - \$1, b, b - \$1, c, c - \$1\}$. Say that Least Bad permits either a or c from this set. If you choose 'Up' at the first node, you are now faced with a choice between a and $c - \$1$. Though $c - \$1 \succ a$, you do not choose $c - \$1$ since you do not take yourself to be choosing relative to $\{a, c - \$1\}$. Rather, the context of evaluation is $\{a, a - \$1, b, b - \$1, c, c - \$1\}$, and $c - \$1$ is not choiceworthy in this set. So, you end up with a.

This style of reasoning results in invulnerability to Money Pumps more generally. In a deterministic decision situation, let O be the set of outcomes available by following some sequence of acts. Then, we can formulate the rule:

> **Retrospective Choice:** If possible, at node n you make a choice that is part of a sequence whose outcome is in $\boldsymbol{c}(O)$.[230]

To see that this avoids Money Pumps, recall that we assume that if $o \in \boldsymbol{c}(O)$, then $o - \$\epsilon \notin \boldsymbol{c}(O)$. So, beginning from the initial choice node, the retrospective chooser never makes a choice that leaves $o - \$\epsilon$ open *without* leaving some choiceworthy element of $\boldsymbol{c}(O)$ open. In particular, they never choose in a way that terminates a sequence with $o - \$\epsilon$, since doing so would rule

[230] Note that this permits a variety of rules to determine $\boldsymbol{c}(O)$—in line with the discussion so far, I use Least Bad. Since every permissible plan results in a permissible outcome (in the set of outcomes possible relative to some plan), Retrospective Choice satisfies Hammond's *Strong Reduction* principle.

out all outcomes in $\boldsymbol{c}(O)$. Retrospective Choice therefore never permits paying a small fee to get something you could have for free.[231]

It is unclear to me that we should prohibit backwards-looking reasoning. It is certainly psychologically plausible that we evaluate options in terms of what we could have had.[232] And agents who do so shield themselves against exploitability.

Moreover, we can motivate Retrospective Choice by examining the relationship between acts in a sequence and the terminal outcome of that sequence. Note that only *terminal nodes* in a decision tree correspond to outcomes—everything else in the sequence is a means to those ends. Now, recall the idea from Briggs (2010b, p. 23) that our goal as rational actors is 'to end up with a future self who is in some sense happy to exist'. In some decision tree, when making a choice at some node, what does it mean to act in order to bring about a future self who 'is in some sense happy to exist'? Presumably that you act to bring about a future self who receives an outcome that is choiceworthy by their lights.[233] But given the context-sensitivity of $\boldsymbol{c}(\cdot)$, it is non-trivial to say when your future self receives something that they judge to be choiceworthy. Plausibly, your future self will think about the outcomes that they could have relative to some sequence of choices that they could have made. So, when your future self ends up with, say, $c - \$1$ in Cantwell's Money Pump will be unhappy with this insofar as they think they could have done better. And they will make this judgement relative to the sequences of choices that were available to them. At each node then, if you act in order to bring about a future self who is happy to exist, and if *they* evaluate outcomes relative to O, then *you* should at each node make a choice that is compatible with an outcome in $c(O)$.

To be clear: I am not claiming that rationality *requires* your future self to judge outcomes relative to O. Perhaps they are 'myopic' and only judge outcomes relative to what they could have had at

[231] More generally, whatever dominance constraints we think govern sequences—that you never choose plans that result in absolutely dominated outcomes, fully dominated outcomes, and so on—so long as $\boldsymbol{c}(\cdot)$ satisfies those constraints synchronically, Retrospective Choice guarantees that you satisfy it diachronically.

[232] Machina (1989), for example, holds a backwards-looking view about how to assess risks (though formally Retrospective Choice says nothing about how present options relate to past risks). Even opponents of cyclic preferences and risk-sensitivity might accept that past choices matter in some cases. For example, Gustafsson (2016, p. 66) allows that past decisions may influence present choiceworthiness for agents with incomplete preferences. He claims (p. 66) that this is plausible in cases of incompleteness, where facts beyond preferences (such as what you did in the past) might fill in 'normative gaps' left by incomplete preferences. Gustafsson further claims that caring about facts beyond preferences is *only* permissible in cases of incompleteness. I think, however, that given the context-sensitivity of choice functions caused by cyclicity, we might similarly look to facts beyond preferences (such as what you did in the past) to inform the option set that our choice function operates on.

[233] I am assuming that your preferences are stable throughout a deterministic decision tree—I set aside important and complicated issues that arise when you and your future self have different preferences over outcomes. Note also that we are dealing with *deterministic* decision trees here, so epistemic differences between you and your future self do not arise here, as they did in the case of decision instability for CDT.

some recent choice node. Or perhaps they are 'circumspect' and only judge outcomes relative to what they could have had relative to feasible sequences of choices (those that they believe they would have had a realistic chance of sticking to). So, I do not think that all agents will be able to follow Retrospective Choice—depending on how your future self evaluates choices, you may not be motivated to stick to a course of action recommended by Retrospective Choice. But no matter—I have already argued that rationality does not *require* immunity to Money Pumps. In the case that at the outset of a decision tree you evaluate outcomes relative to O, but some future self at a terminal node evaluates outcomes relative to $O' \subset O$, you may lack the motivational capacity to stick to the course of action recommended by Retrospective Choice. And in those cases, at the outset of a decision tree you may predict that you end up worse off than you need be (relative to some sequence of choices available at the outset of the decision tree). So, I am not thinking of Retrospective Choice as a universal response to Money Pumps, but as one further tool in our toolkit to minimise exploitability.

A standard objection to rules like Retrospective Choice is that they are *non-consequentialist* (see Hammond 1988a,b)—by letting past facts influence your current decisions, something other than acts' possible outcomes informs your decision-making.[234] And you might think that a theory of instrumental rationality should only evaluate acts based on their possible outcomes.

But we must be careful in saying precisely what it takes for a theory to be 'consequentialist'. At the most basic level, a theory is consequentialist if '[rational] behaviour is explicable merely by its consequences' (Hammond 1988b, p. 503). And Retrospective Choice says that a choice is permissible only if it leads to a permissible consequence (in the relevant context of evaluation, which is O). So, Retrospective Choice does evaluate each act based on its outcomes—nothing other than consequences feature in your evaluations of acts. Retrospective Choice therefore falls under a general characterisation of consequentialism.

If we reject Retrospective Choice on consequentialist grounds, it must be because we have some more specific notion of consequentialism in mind. The idea must be that when evaluating acts at some choice node, only the possible outcomes *of those acts at that choice node* matter. Let O_{A_n} be the set of outcomes that terminate some sequence passing through n. I say that a choice function is *ends directed* if for any outcome set O, the contents of $\boldsymbol{c}(O)$ supervenes on the agent's

[234] See Steele (2010, pp. 470-471) for discussion of Hammond's use of 'consequentialism', particularly as it relates to evaluating plans in normal versus extensive form.

preferences—no change in preferences, no change in permissibility. Then say then that a theory is *narrowly consequentialist* if for each decision situation:

> At choice node n, permissible choices are those that leave open an element in $\boldsymbol{c}(O_{A_n})$, and $\boldsymbol{c}(\cdot)$ is ends directed.

A narrowly consequentialist theory says that when you look at the choices in front of you, you consider only the outcomes that those choices might yield. If you do not choose something at least *compatible* with bringing about an optimal outcome in that set, then it is impermissible. This captures one strong sense in which the goal of rationality is to make choices that bring about optimal consequences.

Retrospective Choice is not narrowly consequentialist. For example, holding fixed your preferences in Cantwell's Money Pump, consider a choice at the node (2,Up). At this node, $c - \$1$ is impermissible since c is available elsewhere in the decision situation. However, were c not available elsewhere, $c - \$1$ might be permissible by the lights of Least Bad. So, what you may do at this node depends on more than the outcomes available from (2,Up). And this makes sense—the retrospective chooser cares about what they could have had in a situation.

I think that the defender of cyclic preferences can reject narrow consequentialism. Given cyclic preferences, context of evaluation matters greatly—your evaluation of each outcome depends crucially on what else is available. So, while we might think that only consequences matter when choosing, we might still ask how you should determine the evaluative status of those consequences. I can consistently maintain 'Each choice in A_n is evaluated only by its possible outcomes' and 'To evaluate those possible outcomes, it matters what I could have had relative to some past time'. (After all, at the end of a decision tree you might judge the outcome you receive relative to what else you *could* have had.) So, a broad commitment to consequentialism will allow that context of evaluation is influenced by a range of factors beyond which outcomes are available now. Recall that in a decision situation N is the set of choice nodes, O is the set of possible outcomes, and let $\wp(O)$ be the power set of O. Say then that a theory is *broadly consequentialist* if in each decision situation:

> The theory describes a map $N \to \wp(O)$, which maps each n to a *context of evaluation* O_n, and permissible choices in A_n are those that leave open an outcome in $\boldsymbol{c}(O_n)$, and $\boldsymbol{c}(\cdot)$ is ends directed.

Essentially, a broadly consequentialist theory is one that at each node tells an agent with fixed preferences to pursue optimal consequences (relative to some context of evaluation). If at node n you do not choose something at least *compatible* with realising an optimal outcome in O_n, then you behave irrationally. So again, the goal of rationality is to make choices that bring about optimal consequences (though we are more expansive in stipulating what it takes to be an optimal consequence).[235]

Retrospective Choice plus Least Bad is broadly consequentialist. Firstly, Least Bad makes $\boldsymbol{c}(\cdot)$ ends directed—since Φ is an SSB function, it is uniquely determined by an agent's preferences, so there can be no change in the Least Bad element in O without a change in the agent's preferences. At each node n, Retrospective Choice defines a simple map, n to O, and it then says that permissible choices are just those that are compatible with an element of $\boldsymbol{c}(O)$. In this way, Retrospective Choice's recommendations are based only a concern for promoting appropriate consequences (relative to the correct context of evaluation, O).[236,237]

Finally, say that a theory is *non-consequentialist* if it is not broadly consequentialist. An example of a non-consequentialist theory would be:

> **Promise-Keeping:** At node n, a is impermissible only if you promised not to do a at some previous time.

To see that this rule is not broadly consequentialist, take an agent who initially goes 'Up' in Cantwell's Money Pump, giving them a choice between a and $c - \$1$. If Promise-Keeping is broadly consequentialist, then this choice node gets mapped to some context of evaluation, O_{Up},

[235] A broadly consequentialist theory need not satisfy Dynamic Separability. As I have said, that is a bullet that I am prepared to bite—given the context-sensitivity of choice functions, we can allow that you reason in light of what you could have had.

[236] Gustafsson (2016, p. 65) wants to uphold the following consequentialist constraint, which is discussed by Peterson (2013, p. 135):

> **Normative Supervenience:** The normative statuses of the available alternatives in a situation are determined by the evaluative ranking of those alternatives.

Provided we interpret 'situation' and 'evaluative ranking' appropriately, Retrospective Choice plus Least Bad satisfies this supervenience principle. The normative statuses of the available alternatives (at a choice node) are determined by *their Least Bad scores in* O. Least Bad scores are a kind of evaluative ranking—they order outcomes in an option set from least to most bad. So, what you may do at each choice node depends only on (i) the available alternatives at that node, and (ii) the evaluative rankings of each of those alternatives. So, there is a plausible reading of Normative Supervenience that allows for Retrospective Choice. Of course, Resolute Choice plus Least Bad is incompatible with a reading of Normative Supervenience on which evaluative ranking of options are dynamically separable (since Least Bad scores in O might depend on facts about what you could have had in the past).

[237] Hammond (1988, p. 518) argues that backwards-looking considerations should be built into descriptions of outcomes: 'Perhaps history matters. But if it does, it is a relevant consequence'. But the discussion so far indicates a way that history might matter that is distinct from it being part of a consequence. History might influence the context relative to which you evaluate outcomes, and contra Hammond, this is not the same as having preferences that are influenced by historical concerns.

such that the agent makes a choice that leaves open an element of $\boldsymbol{c}(O_{Up})$. If Promise-Keeping is broadly consequentialist, we can hold the contents of O_{Up} fixed while varying whether the agent made a past promise—if not, then the theory would not define a map $N \rightarrow \wp(O)$. So, if the agent made no past promises, then by Promise-Keeping the agent may choose a and it must be that $a \in \boldsymbol{c}(O_{Up})$. But if the agent made a promise not to do a, then they may not choose a and it must be that $a \notin \boldsymbol{c}(O_{Up})$. So, Promise-Keeping is not broadly consequentialist—the goal of this rule is not to promote optimal consequences in some context of evaluation settled by the structure of the decision tree. In general, a rule is not consequentialist if it lets permissibility vary with your past commitments, intentions, plans, and so on.[238]

Backwards-looking reasoning is consequentialist enough for me. Nothing over and above outcomes matters when evaluating acts. It is true that you cannot work out what to do by considering just the outcomes possible *from the time* you make your decision. But if we accept that choice functions are highly context-sensitive, then it makes sense that facts about the past can, by way of setting the context of evaluation, affect your practical reasoning. That practical reasoning is, however, still directed towards bringing about appropriate consequences.

It is important to distinguish Retrospective Choice from views that say that past intentions are reason-giving. Retrospective Choice never mentions what you actually did, planned, committed to, or intended in the past. So, Retrospective Choice is compatible with, say, Broome's (2001) view that what you previously intended provides no reason(s) to choose some options over others. Indeed, say that at some past time you intended to turn down b. Retrospective Choice does not say that you have any additional reason to (dis)prefer b over a or anything else. The fact that b was possible for you at some time means that comparisons with b are now relevant for the purposes of practical deliberation. But that is just to say that you structure your decision-making in a certain way, not that what you intended has any bearing on your essentially comparative reasons or judgements of betterness.[239]

[238] This means that some versions of Resolute Choice—on which what you may do at a node depends on whether you made a resolution in the past—are non-consequentialist on my view. Note that Retrospective Choice is neutral on whether you can have a preference for carrying out past commitments (cf. McClennen 1988, Sobel 1988a). Nonetheless, Retrospective Choice allows for dynamic unification *without* your having a preference for carrying out plans. The Humean in me resists that we must have certain goals (say, an intrinsic preference for carrying out plans) to achieve dynamic unification. So, I take it to be an advantage of Retrospective Choice that it allows you to pursue unified plans without placing any restrictions on your preferences over outcomes.

[239] This likewise distinguishes Retrospective Choice from, say, the sunk-cost fallacy. The sunk-cost fallacy involves doing something sub-optimal because of costs that you bore in the past. But once we specify O as the appropriate context of evaluation, Retrospective Choice recommends only choiceworthy options in O, regardless of what you did in the past. So there can be no question of choosing something sub-optimal because of a past decision.

So, Retrospective Choice is psychologically plausible, avoids Money Pumps, and is broadly consequentialist. Though it says that permissibility depends on past facts, this is importantly different from controversial views that make permissibility a matter of past decisions, commitments, or intentions. I do not claim that Retrospective Choice is available to all agents. It is an empirical question to what degree we are motivated by counterfactual considerations, so Retrospective Choice may not be the kind of rule that every rational agent can adopt. But it is one further tool that, when available, shields us from exploitation.[240]

6.6 Conclusion: Close Enough for Jazz

I have argued that Least Bad is a reasonable response to the general choice problem. I have not argued that it is the only plausible response, but it satisfies some minimal normative properties, is simple, and vindicates widely held intuitions. This sheds important light on which dominance constraints are genuine normative requirements. Not only should the Non-Dominated Choice constraint go, I have suggested that the Non-Fully Dominated Choice can also be reasonably rejected. But for the fan of dominance, things are not all doom and gloom. Least Bad will recommend non-dominated options when available, and mixed-acts may allow us to perform (or approximate performing) non-dominated options in more situations than is typically acknowledged.

In the dynamic setting, I have argued that there is nothing incoherent in being Money Pumped, nor selecting a dominated sequence of choices. Nonetheless, I have developed two tools here that smooth off some of the hard edges of that position. The existence of reference-act dependent EU functions provides a tool for agents to rework their preferences and so avoid exploitation. And Retrospective Choice provides a broadly consequentialist strategy to maintain cyclic preferences but still avoid dominated sequences. So, while accepting cyclic preferences does force us to reject some widely endorsed dominance principles, we have plenty of tools to minimise how often we violate such principles, as well as the severity of such violations.

[240] Note that backwards-looking reasoning does not rule out forwards-looking reasoning (see, for example, Rabinowicz 1995 who proposes Wise Choice, which combines backward induction with the possibility of Resolute Choice). Rabinowicz focusses on Resolute Choice, not Retrospective Choice—considering hybrid models of Retrospective Choice and forwards-looking strategies is for future work.

Conclusion

What are the correct dominance principles? That is a hard question—in philosophical discussions, dominance principles typically feature as *tools* with which we assess other principles, and it is hard to assess those principles directly.

My approach here has been to investigate dominance principles from various angles, and in each case to explore the conflict(s) that arise between principles. Some of those conflicts are enmeshed in well-established debates (e.g., the conflict between causal and non-causal versions of State-wise Dominance). Others are relatively underexplored by philosophers (e.g., Betweenness and Decomposability as constraints on preferences, and dominance in the context of cyclic preferences).

Stepping back from individual debates, we can see that a handful of interrelated themes crop up repeatedly. One is the role of *ex post* regret—when, if ever, is it rational to do something that your better-informed self prefers you not to do? Another is the relationship between prospects and motivation—do probability distributions over outcomes exhaust everything we care about? And another is how to behave when we lack a 'stable standpoint'—what does rationality require when our views about what is preferable change over the course of a sequence of decisions? I do not claim to have settled any of these questions definitively. Rather, I hope to have shown some of the contour (or fault) lines that characterise dominance across a range of areas.

So, where does that leave us? I began by noting that philosophers talk about dominance with respect to *preferences*, *choices*, and *plans*. While recognising that much more could be said about each issue, I have developed a view—Least Bad—that satisfies the following:

- Constraints on preference: uphold Mixture Dominance (but reject State-wise Dominance of all varieties).
- Constraints on choice: uphold the Sufficiency of Non-Dominated Choices (but reject even the Non-Fully Dominated Choice constraint).
- Constraints on plans: reject all forms of Non-Dominated Plan constraint.

This is a far more minimal view of dominance than many would accept. All we can really say is that (i) randomising over worse (better) options does not make things better (worse), and (ii) if something is as good as everything else, then it is an appropriate means to your ends.

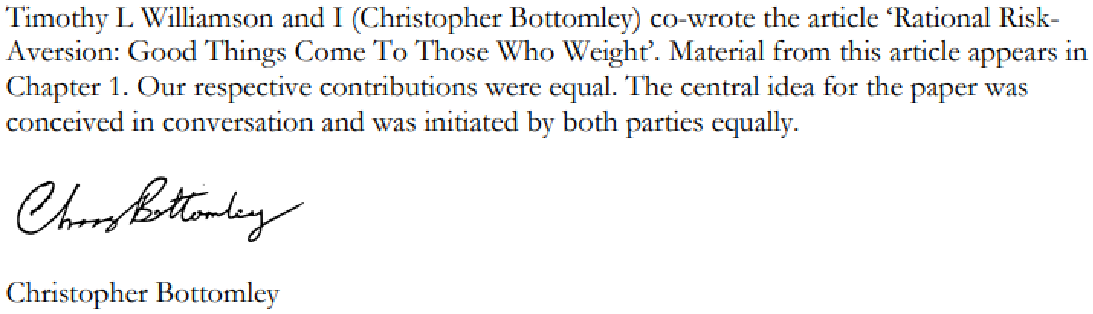
If we accept this picture, do we thereby relegate most familiar dominance constraints to the status of uninformative or unhelpful? I do not think so. While those principles admit of *exceptions*, they may still play a role in simplifying and analysing reasoning. For example, if we adopt SSB and reject State-wise Dominance as a general constraint, we can note that SSB-maximisers (even those with cyclic preferences) respect State-wise Dominance in special cases where the set of states is appropriately structured (see Fishburn 1989, p. 192). And even if we reject the Non-Fully Dominated Choices constraint, agents will choose non-dominated options when mixed-acts are available, or at least *approximate* doing so. Moreover, we know that SSB respects First-Order Stochastic Dominance when outcomes are transitively ordered (see Fishburn 1984b, Theorem 8). So, if outcomes are, say, transitively ordered sums of money (as is often useful in economic applications), then we can apply a standard dominance tool to predict and explain behaviour.

Indeed, we can uphold State-wise Dominance in the special case that preferences satisfy Transitivity. Of course, reasonable agents might disagree about the precise details of State-wise Dominance—you might adopt a broadly causal formulation, while I might reject it. But the well-known 'Tickle Defence' (see Eells 1982, Price 1986, Briggs 2010b Section 7) indicates that Causal and Evidential Decision Theories will often coincide—the distinction between various kinds of State-wise Dominance will therefore often not matter. So, in many cases we can sidestep complicated issues about how precisely to define act-state independence.

This all indicates that we might think of dominance principles as *domain-specific* tools. Some familiar dominance principles apply only given acyclic preferences, others apply only when the set of states is appropriately structured, and others apply only in the case that agents have access to mixed-acts. We can indicate when agents are not living up to their own standards, and we know that agents will recognise that in a wide range of circumstances they *should* avoid dominating options.

So, dominance constraints might admit of exceptions, but they are still useful. They allow us to simplify decisions, predict and explain behaviour, and criticise agents for doing worse than they could have. The fact that we can only do so under the right conditions should not be surprising—all tools have domains of applicability. So, while many dominance principles admit of exceptions, they may still make our lives easier both as decision-makers and decision-theorists looking to understand and evaluate choices.

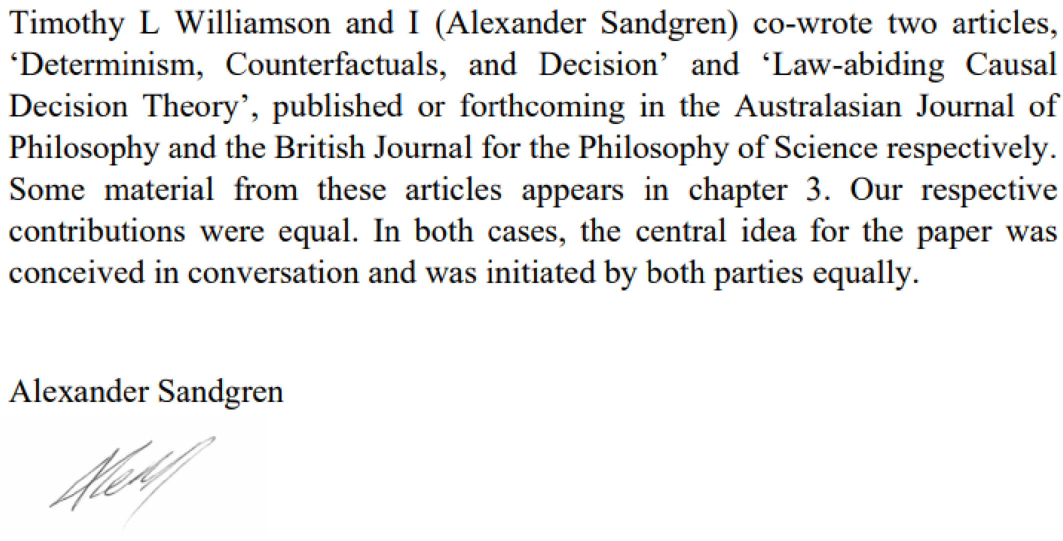
Appendix: Signed Statements of Co-Authorship



Timothy L Williamson and I (Christopher Bottomley) co-wrote the article 'Rational Risk-Aversion: Good Things Come To Those Who Weight'. Material from this article appears in Chapter 1. Our respective contributions were equal. The central idea for the paper was conceived in conversation and was initiated by both parties equally.

Chris Bottomley

Christopher Bottomley



Timothy L Williamson and I (Alexander Sandgren) co-wrote two articles, 'Determinism, Counterfactuals, and Decision' and 'Law-abiding Causal Decision Theory', published or forthcoming in the Australasian Journal of Philosophy and the British Journal for the Philosophy of Science respectively. Some material from these articles appears in chapter 3. Our respective contributions were equal. In both cases, the central idea for the paper was conceived in conversation and was initiated by both parties equally.

Alexander Sandgren

Bibliography

Adams, E & Rosenkrantz, R 1980, 'Applying the Jeffrey Decision Model to Rational Betting and Information Acquisition', *Theory and Decision*, 12, pp. 1-20.

Ahmed, A 2012, 'Push the Button', *Philosophy of Science*, vol. 79, no. 3, pp. 386-395.

Ahmed, A 2013, 'Causal Decision Theory: A Counterexample', *Philosophical Review*, vol. 122, no. 2, pp. 289-306.

Ahmed, A 2014a, 'Causal Decision Theory and the Fixity of the Past', *The British Journal for the Philosophy of Science*, vol. 65. No. 4, pp. 665-685.

Ahmed, A 2014b, *Evidence, Decision and Causality*, Cambridge University Press, Cambridge.

Ahmed, A 2015, 'Infallibility in the Newcomb Problem', *Erkenntnis*, vol. 80, no. 2, pp. 261-273.

Ahmed, A 2016, 'Review of Risk and Rationality', *The British Journal for the Philosophy of Science Review of Books.*

Ahmed, A 2017, 'Exploiting Cyclic Preferences', *Mind*, vol. 126, no. 504, pp. 975-1022.

Ahmed, A 2020, 'Equal Opportunities in Newcomb's Problem and Elsewhere', *Mind*, vol. 129, no. 515, pp. 567-886.

Ahmed, A & Price, H 2012, 'Arntzenius on "Why Ain'cha Rich?"', *Erkenntnis*, vol. 77, no. 1, pp. 15-30.

Ahmed, A & Spencer, J 2020, 'Objective Value Is Always Newcombizable', *Mind*, vol. 129, no. 516, pp. 1157-1192.

Allais, M 1953, 'Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine', *Econometrica*, vol. 21, no. 4, pp. 503-546.

Anand, P 1993, 'The Philosophy of Intransitive Preference', *The Economic Journal*, vol. 103, no. 417, pp. 337-346.

Anand, P 2009, 'Rationality and Intransitive Preference: Foundations for the Modern View', in P Anand, P Pattanaik & C Puppe (eds), *The Handbook of Rational and Social Choice*, Oxford University Press, Oxford.

Andreou, C 2016, 'Cashing Out the Money-Pump Argument', *Philosophical Studies*, vol. 173, pp. 1451-1455.

Anscombe, F & Aumann, R 1963, 'A Definition of Subjective Probability', *Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 199-205.

Armendt, B 1980, 'Is There a Dutch Book Argument for Probability Kinematics?', *Philosophy of Science*, vol. 47, pp. 583-588.

Armendt, B 1992, 'Dutch Strategies for Diachronic Rules: When Believers See the Sure Loss Coming', in D Hull, M Forbes & K Okruhlik (eds), *PSA 1992*, Philosophy of Science Association, East Lansing.

Armendt, B 2014, 'On Risk and Rationality', *Erkenntnis*, vol. 79, no. S6, pp. 1-9.

Armendt, B 2019, 'Causal Decision Theory and Decision Instability', *The Journal of Philosophy*, vol. 116, no. 5, pp. 263-277.

Arntzenius, F 2008, 'No Regrets, Or: Edith Piaf Revamps Decision Theory', *Erkenntnis*, vol. 68, no. 2, pp. 277-297.

Arntzenius, F, Elga, A & Hawthorne, J 2004, 'Bayesianism, Infinite Decisions, and Binding', *Mind*, vol. 113, no. 450, pp. 251-283.

Bader, R 2018, 'Stochastic Dominance and Opaque Sweetening', *Australasian Journal of Philosophy*, vol. 96, no. 3, pp. 498-507.

Bales, A 2016, 'The Pauper's Problem: Chance, Foreknowledge, and Causal Decision Theory', *Philosophical Studies*, vol. 173, no. 6, pp. 1497-1516.

Bales, A 2018a, 'Decision-Theoretic Pluralism', *Philosophical Quarterly*, vol. 68, no. 273, pp. 801-818.

Bales, A 2018b, 'Richness and Rationality: Causal Decision Theory and the WAR Argument', *Synthese*, vol. 195, no. 1, pp. 259-267.

Bales, A 2018c, 'Indeterminate Permissibility and Choiceworthy Options', *Philosophical Studies*, vol. 175, no. 7, pp. 1693-1702.

Bales, A 2020, 'Intentions and Instability: A Defence of Causal Decision Theory', *Philosophical Studies*, vol. 177, no. 3, pp. 783-804.

Baron, J 2008, *Thinking and Deciding,* 4th edn, Cambridge University Press, Cambridge.

Bartha, P 2007, 'Taking Stock of Infinite Value: Pascal's Wager and Relative Utilities', *Synthese*, vol. 154, no. 1, pp. 5-52.

Bartha, P, Barker, J & Hájek, A 2014, 'Satan, Saint Peter and Saint Petersburg: Decision Theory and Discontinuity at Infinity', *Synthese*, vol. 191, no. 4, pp. 629-660.

Bernstein, S 2016, 'Omission Impossible', *Philosophical Studies*, vol. 173, no. 10, pp. 2575-2589.

Blessenohl, S 2020, 'Risk Attitudes and Social Choice', *Ethics*, vol. 130, no. 4, pp. 485-513.

Bottomley, C & Williamson, TL 2021, 'Reasonable Risk-Aversion: Good Things Come to those who Weight', manuscript.

Bradley, R 2017, *Decision Theory with a Human Face*, Cambridge University Press, Cambridge.

Bratman, M 1987, *Intentions, Plans and Practical Reason*, Harvard University Press, Cambridge.

Briggs, R 2010a, 'Putting and Value on Beauty', in TS Gendler & J Hawthorne (eds), *Oxford Studies in Epistemology: Volume 3*, Oxford University Press, Oxford, pp. 3-34.

Briggs, R 2010b, 'Decision-Theoretic Paradoxes as Voting Paradoxes', *Philosophical Review*, vol. 119, no. 1, pp. 1-30.

Briggs, R 2015, 'Costs of Abandoning the Sure-Thing Principle', *Canadian Journal of Philosophy*, vol. 45, no. 5, pp. 827-840.

Briggs, R 2019, 'Normative Theories of Rational Choice: Expected Utility', in *The Stanford Encyclopedia of Philosophy*. Available at: <https://plato.stanford.edu/entries/rationality-normative-utility/#LonRunArg>

Briggs, R & Pettigrew, R 2020, 'An Accuracy-Dominance Argument for Conditionalization', *Noûs*, vol. 54, no. 1, pp. 162-181.

Broome, J 1991, *Weighing Goods: Equality, Uncertainty, and Time*, Blackwell, Oxford.

Broome, J 2001, 'Are Intentions Reasons? And How Should We Cope with Incommensurable Values', in *Practical Rationality and Preference: Essays for David Gauthier*, C Morris and A Ripstein (eds), Cambridge University Press, Cambridge, pp. 98-120.

Broome, J 2013, *Rationality through Reasoning*, Wiley-Blackwell, West Sussex.

Brown, C 2020, 'Is Close Enough Good Enough?' *Economics and Philosophy*, vol. 36, no. 1, pp. 29-59.

Buchak, L 2013, *Risk and Rationality*, Oxford University Press, Oxford.

Buchak, L 2014, 'Risk and Tradeoffs', *Erkenntnis*, vol. 79, no. S6, pp. 1091-1117.

Camerer, C 1989, 'An Experimental Test of Several Generalized Utility Theories', *Journal of Risk and Uncertainty*, vol. 2, no. 1, pp. 61-104.

Camerer, C & Ho, T 1994, 'Violations of the Betweenness Axiom and Nonlinearity in Probability', *Journal of Risk and Uncertainty*, vol. 8, no. 2, pp. 167-196.

Cantwell, J 2003, 'On the Foundations of Pragmatic Arguments', *The Journal of Philosophy*, vol. 100, no. 8, pp. 383-402.

Chew, SH 1983, 'A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox', *Econometrica*, vol. 51, no. 4, pp. 1065-1092.

Chew, SH 1989, 'Axiomatic Utility Theories with the Betweenness Property', *Annals of Operations Research*, vol. 19, pp. 273-298.

Christensen, D 1991, 'Clever Bookies and Coherent Beliefs', *Philosophical Review*, vol. 100, no. 2, pp. 229-247.

Cohen, M 1995, 'Risk-Aversion Concepts in Expected- and Non-Expected Utility Models', *The Geneva Papers on Risk and Insurance Theory*, vol. 20, no. 1, pp. 73-91.

Colyvan, M 2008, 'Relative Expectation Theory', *Journal of Philosophy*, vol. 105, no. 1, pp. 37-44.

Davidson, D, McKinsey, JCC & Suppes, P 1955, 'Outlines of a Formal Theory of Value, I', *Philosophy of Science*, vol. 22, no. 2, pp. 140-160.

Dekel, E 1986, 'An Axiomatic Characterization of Preferences Under Uncertainty: Weakening the Independence Axiom', *Journal of Economic Theory*, vol. 40, no. 2, pp. 304-318.

Diamond, P 1967, 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: A Comment', *Journal of Political Economy,* vol. 75, no. 5, pp. 765-766.

Diecidue, E, Schmidt, U & Zank, H 2009, 'Parametric Weighting Functions', *Journal of Economic Theory*, vol. 144, no. 3, pp. 1102-1118.

Doody, R 2016, 'Doing Your Best (While Making Do with Less): The Actual Value Conception of Instrumental Rationality', PhD Dissertation, Massacheusetts Institute of Technology, Cambridge, Massachusetts. Available at: <https://dspace.mit.edu/handle/1721.1/107095>

Doody, R 2019a, 'Parity, Prospects, and Predominance', *Philosophical Studies*, vol. 176, no. 4, pp. 1077-1095.

Doody, R 2019b, 'Opaque Sweetening and Transitivity', *Australasian Journal of Philosophy*, vol. 97, no. 3, pp. 559-571.

Dorr, C 2016, 'Against Counterfactual Miracles', *Philosophical Review*, vol. 125, no. 2, pp. 241-286.

Dougherty, T 2014, 'A Deluxe Money Pump', *Thought: A Journal of Philosophy*, vol. 3, no. 1, pp. 21-29.

Easwaran, K 2014a, 'Principal Values and Weak Expectations', *Mind*, vol. 123, no. 490, pp. 517-531.

Easwaran, K 2014b, 'Decision Theory without Representation Theorems', *Philosophers' Imprint*, vol. 14, no. 27, pp. 1-30.

Easwaran, K Forthcoming, 'A Classification of Newcomb Problems and Decision Theories', *Synthese*.

Eells, E 1982, *Rational Decision and Causality*, Cambridge University Press, Cambridge.

Egan, A 2007, 'Some Counterexamples to Causal Decision Theory', *Philosophical Review*, vol. 116, no. 1, pp. 93-114.

Elga, A 2010, 'Subjective Probabilities should be Sharp', *Philosophers' Imprint*, vol. 10, no. 5, pp. 1-11.

Eriksson, L & Rabinowicz, R 2013, 'The Interference Problem for the Betting Interpretation of Degrees of Belief', *Synthese*, vol. 190, no. 5, pp. 809-830.

Finucane, ML, Alhakami, A, Slovic, P & Johnson, SM 2000, 'The Affect Heuristic in Judgements of Risks and Benefits', *Journal of Behavioural Decision Making*, vol. 13, no. 1, pp. 1-17.

Fishburn, P 1978, 'Stochastic Dominance without Transitive Preferences', *Management Science*, vol. 24, no. 12, pp. 1268-1277.

Fishburn, P 1981, 'Subjective Expected Utility: A Review of Normative Theories', *Theory and Decision*, vol. 13, pp. 139-199.

Fishburn, P 1982, 'Nontransitive Measurable Utility', *Journal of Mathematical Psychology*, vol. 26, no. 1, pp. 31-67.

Fishburn, P 1984a, 'SSB Utility Theory: An Economic Perspective', *Mathematical Social Sciences*, vol. 8, no. 1, pp. 63-94.

Fishburn, P 1984b, 'Dominance in SSB Utility Theory', *Journal of Economic Theory*, vol. 34, no. 1, pp. 130-148.

Fishburn, P 1988, *Nonlinear Preference and Utility Theory*, The Johns Hopkins University Press, Baltimore.

Fishburn, P 1989a, 'Stochastic Dominance in Nonlinear Utility Theory', in TB Fomby & TK Seo (eds), *Studies in the Economics of Uncertainty*, Springer, New York, pp. 3-20.

Fishburn, P 1989b, 'Non-transitive Measurable Utility for Decision under Uncertainty', *Journal of Mathematical Economics*, vol. 18, no. 2, pp. 187-207.

Fishburn, P 1990, 'Skew Symmetric Additive Utility with Finite States', *Mathematical Social Sciences*, vol. 19, no. 2, pp. 103-115.

Fishchoff, B, Slovic, P, Lichtenstein, S, Read, S & Combs, B 1978, 'How Safe is Safe Enough? A Psychometric Study of Attitudes Towards Technological Risks and Benefits', *Policy Sciences*, vol. 9, no. 2, pp. 127-152.

Gallow, JD 2020, 'The Causal Decision Theorist's Guide to Managing the News', *The Journal of Philosophy*, vol. 117, no. 3, p. 117-149.

Gallow, JD Forthcoming, 'Escaping the Cycle', *Mind.*

Gauthier, D 1994, 'Assure and Threaten', *Ethics*, vol. 104, no. 4, pp. 690-716.

Gibbard, A & Harper, W 1978, 'Counterfactuals and Two Kinds of Expected Utility', in C Hooker, J Leach and E McClennen (eds), *Foundations and Applications of Decision Theory: Volume 1*, Reidel, Boston.

Gibbard, A 1986, 'Characterisation of Decision Matrices that Yield Instrumental Expected Utility', in L Daboni, A Montesano and M Lines (eds), *Recent Developments in the Foundations of Utility and Risk Theory*, D. Reidel, Boston, pp. 139-148.

Good, IJ 1967, 'On the Principle of Total Evidence', *The British Journal for the Philosophy of Science*, vol. 17, no. 4, pp. 319-321.

Goodman, J 2015, 'Knowledge, Counterfactuals, and Determinism', *Philosophical Studies*, vol. 172, no. 9, pp. 2275-2278.

Grant, S, Kajii, A & Polak, B 1992, 'Many Good Choice Axioms: When Can Many-Good Lotteries Be Treated as Money Lotteries?', *Journal of Economic Theory*, vol. 56, no. 2, pp. 313-337.

Grant, S 1995, 'Probabilistic Sophistication without Monotonicity: or How Machina's Mom May Also be Probabilistically Sophisticated', *Econometrica*, vol. 63, no. 1, pp. 159-189.

Grant, S, Kajii, A & Polak, B 2000, 'Decomposable Choice under Uncertainty', *Journal of Economic Theory*, vol. 92, no. 2, pp. 169-197.

Grant, S, Özsoy, H & Polak, P 2008, 'Probabilistic Sophistication and Stochastic Monotonicity in the Savage Framework', *Mathematical Social Sciences*, vol. 55, no. 3, pp. 371-380.

Gul, F 1991, 'A Theory of Disappointment Aversion', *Econometrica*, vol. 59, no. 3, pp. 667-686.

Gul, F 1992, 'Savage's Theorem with a Finite Number of States', *Journal of Economic Theory*, vol. 57, no. 1, pp. 99-110.

Gul, F & Lantto, O 1990, 'Betweenness Satisfying Preferences and Dynamic Choice', *Journal of Economic Theory*, vol. 52, no. 1, pp. 162-177.

Gustafsson, J 2011, 'A Note in Defence of Ratificationism', *Erkenntnis*, vol. 75, no. 1, pp. 147-150.

Gustafsson, J 2013, 'The Irrelevance of the Diachronic Money-Pump Argument for Acyclicity', *The Journal of Philosophy*, vol. 110, no. 8, pp. 460-464.

Gustafsson, J 2016, 'Money Pumps, Incompleteness, and Indeterminacy', *Philosophy and Phenomenological Research*, vol. 92, no. 1, pp. 60-72.

Gustafsson, J Forthcoming, '*Ex Ante* Prioritarianism Violates *Ex Ante* Pareto', *Utilitas.*

Hájek, A 2008, 'Arguments for—or against—Probabilism?', *The British Journal for the Philosophy of Science*, vol. 59, no. 4, pp. 793-819.

Hájek, A 2014a, 'Unexpected Expectations', *Mind*, vol. 123, no. 490, pp. 503-516.

Hájek, A 2014b, 'A Chancy "Magic Trick"', in A Wilson (ed.), *Chance and Temporal Asymmetry*, Oxford University Press, Oxford, pp. 100-110.

Hájek, A 2016, 'Deliberation Welcomes Prediction', *Episteme*, vol. 13, no. 4, pp. 507-528.

Hájek, A Forthcoming[a], 'Risky Business', *Philosophical Perspectives.*

Hájek, A Forthcoming[b], 'Contra Counterfactism', *Synthese.*

Hájek, A & Nover, H 2004, 'Vexing Expectations', *Mind*, vol. 113, no. 450, pp. 237-249.

Hájek, A & Nover, H 2006, 'Perplexing Expectations', *Mind*, vol. 115, no. 459, pp. 703-720.

Hájek, A & Smithson, M 2012, 'Rationality and Indeterminate Probabilities', *Synthese*, vol. 187, no. 1, pp. 33-48.

Hammond, P 1976, 'Changing Tastes and Coherent Dynamic Choice', *Review of Economic Studies*, vol. 43, no. 1, pp. 159-173.

Hammond, P 1977, 'Dynamic Restrictions on Metastatic Choice', *Economica*, vol. 44, no. 176, pp. 337-350.

Hammond, P 1988a, 'Consequentialist Foundations for Expected Utility', *Theory and Decision*, vol. 25, pp. 25-78.

Hammond, P 1988b, 'Consequentialism and the Independence Axiom', in in B Munier (ed.), *Risk, Decision and Rationality*, D. Reidel, Holland, pp. 537-542.

Hansson, SO & Grüne-Yanoff, T 2017, 'Preferences', in *The Stanford Encyclopedia of Philosophy*. Accessible at: <https://plato.stanford.edu/archives/sum2018/entries/preferences/>

Hare, C 2010, 'Take the Sugar', *Analysis*, vol. 70, no. 2, pp. 237-247.

Hare, C 2016, 'Should we Wish Well to All?', *Philosophical Review*, vol. 125, no. 4, pp. 451-472.

Hare, C & Hedden, B 2016, 'Self-Reinforcing and Self-Frustrating Decisions', *Noûs*, vol. 50, no. 3, pp. 604-628.

Harper, W 1986, 'Mixed Strategies and Ratifiability in Causal Decision Theory', *Erkenntnis*, vol. 24, no. 1, pp. 25-36.

Harsanyi, J 1977, 'On the Rationale of the Bayesian Approach: Comments on Proferssor Watkins's Paper', in R Butts and J Hintikka (eds), *Foundational Problems in the Special Sciences*, D. Reidel, Holland, pp. 381-392.

Hausman, D 2011, *Preference, Value, Choice, and Welfare*, Cambridge University Press, Cambridge.

Hedden, B 2012, 'Options and the Subjective Ought', *Philosophical Studies*, vol. 158, no. 2, pp.343-360.

Hedden, B 2015, 'Options and Diachronic Tragedy', *Philosophy and Phenomenological Research*', vol. 90, no. 2, pp. 423-451.

Hedden, B 2020, 'Consequentialism and Collective Action', *Ethics*, vol. 130, no. 4, pp. 530-554.

Hedden, B Manuscript, 'Counterfactual Decision Theory'.

Herlitz, A 2020, 'Non-transitive Better than Relations and Rational Choice', *Philosophia*, vol. 48, no. 1, pp. 178-189.

Horgan, T 1981, 'Counterfactuals and Newcomb's Problem', *The Journal of Philosophy*, vol. 78, no. 6, pp. 331-356.

Horgan, T 2017, *Essays on Paradoxes*, Oxford University Press, Oxford.

Huemer, M 2008, 'In Defence of Repugnance', *Mind*, vol. 117, no. 468, pp. 899-933.

Hutteger, S & Rothfus, G Forthcoming, 'Bradley Conditionals and Dynamic Choice', *Synthese.*

Isaacs, Y 2016, 'Probabilities Cannot Be Rationally Neglected', *Mind*, vol. 125, no. 299, pp. 759-762.

Jackson, F 1991, 'Decision-Theoretic Consequentialism and the Nearest and Dearest Objection', *Ethics*, vol. 101, no. 3, pp. 461-482.

Jeffrey, R 1983, *The Logic of Decision*, 2nd edn, University of Chicago Press, Chicago.

Jeffrey, R 2004, *Subjective Probability: The Real Thing*, Cambridge University Press, Cambridge.

Joyce, J 1999, *The Foundations of Causal Decision Theory*, Cambridge University Press, Cambridge.

Joyce, J 2012, 'Regret and Instability in Causal Decision Theory', *Synthese*, vol. 187, no. 1, pp. 123-145.

Joyce, J 2016, 'Review of *Evidence, Decision and Causality* by Arif Ahmed', *The Journal of Philosophy*, vol. 113, no. 5, pp. 224-232.

Kahneman, D & Tversky, K 1979, 'Prospect Theory: An Analysis of Decision Under Risk', *Econometrica*, vol. 72, no. 2, pp. 263-291.

Karni, E, Schmeidler, D & Vind, K 1983, 'On State Dependent Preferences and Subjective Probabilities', *Econometrica*, vol. 51, no. 4, pp. 1021-1031.

Karni, E & Schmeidler, D 2016, 'An Expected Utility Theory for State-Dependent Preferences', *Theory and Decision*, vol. 81, pp. 467-478.

Keller, C, Siegrist, M & Gutscher, H 2006, 'The Role of the Affect and Availability Heuristics in Risk Communication', *Risk Analysis*, vol. 26, no. 3, pp. 631-639.

Levi, I 2002, 'Money Pumps and Diachronic Dutch Books', *Philosophy of Science*, vol. 69, no. S3, pp.235-247.

Lewis, D 1973, *Counterfactuals*, Harvard University Press, Cambridge.

Lewis, D 1979, 'Counterfactual Dependence and Time's Arrow', *Noûs*, vol. 13, no. 4, pp. 455-476.

Lewis, D 1981a, 'Causal Decision Theory, *The Australasian Journal of Philosophy*, vol. 59, no. 1, pp. 5-30.

Lewis, D 1981b, 'Are We Free to Break the Laws?', *Theoria*, vol. 47, no. 3, pp. 113-121.

List, C 2014, 'Free Will, Determinism, and the Possibility of Doing Otherwise', *Noûs*, vol. 48, no. 1, pp. 156-178.

Loomes, G & Sugden, R 1982, 'Regret Theory: An Alternative Theory of Rational Choice under Uncertainty', *The Economic Journal*, vol. 92, no. 368, pp. 805-824.

Loomes, G & Sugden, R 1986, 'Disappointment and Dynamic Consistency in Choice under Uncertainty', *Review of Economic Studies*, vol. 53, no. 2, pp. 271-282.

Loomes, G & Sugden, R 1987, 'Some Implications of a More General Form of Regret Theory', *Journal of Economic Theory*, vol. 41, no. 2, pp. 270-287.

Lundgren, B & Stefánsson, HO 2020, 'Against the *De Minimis* Principle', *Risk Analysis*, vol. 40, no. 5, pp. 908-914.

MacAskill, W 2016, 'Normative Uncertainty as a Voting Problem', *Mind*, vol. 125, no. 500, pp. 967-1004.

Machina, M 1989, 'Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty', *Journal of Economic Literature*, vol. 27, no. 4, pp. 1622-1668.

Machina, M & Schmeidler, D 1992, 'A More Robust Definition of Subjective Probability', *Econometrica*, vol. 60, no. 4, pp. 745-780.

Marschak, J 1950, 'Rational Behaviour, Uncertain Prospects, and Measurable Utility', *Econometrica*, vol. 18, no. 2, pp. 111-141.

Mann, K Forthcoming, 'Relevance and Non-Binary Choices', *Ethics.*

McClennen, E 1988, 'Dynamic Choice and Rationality', in B Munier (ed.), *Risk, Decision and Rationality*, D. Reidel, Holland, pp. 517-536.

McClennen, E 1990, *Rationality and Dynamic Choice*, Cambridge University Press, New York.

Meacham, C 2010, 'Binding and its Consequences', *Philosophical Studies*, vol. 149, no. 1, pp. 49-71.

Meacham, C 2020, 'Difference Minimizing Utility Theory', *Ergo*, vol. 6, no. 35, pp. 999-1034.

Meacham, C & Weisberg, J 2011, 'Representation Theorems and the Foundations of Decision Theory', *Australasian Journal of Philosophy*, vol. 89, no. 4, pp. 641-663.

MacIntosh, D 2010, 'Intransitive Preferences, Vagueness, and the Structure of Procrastination', in C Andreou and M White (eds), *The Thief of Time: Philosophical Essays on Procrastination*, Oxford University Press, Oxford, pp. 68-86.

Morris, M, Sim, D & Girotto, V 1998, 'Distinguishing Sources of Cooperation in the One-Round Prisoner's Dilemma: Evidence for Cooperative Decisions Based on the Illusion of Control', *Journal of Experimental and Social Psychology*, vol. 34, no. 5, pp. 494-512.

Muñoz, D & Spencer, J 2021, 'Knowledge of Objective 'Oughts': Monotonicity and the New Miner's Puzzle', *Philosophy and Phenomenological Research*, vol. 103, no. 1, pp. 77-91.

Nebel, J 2018, 'The Good, the Bad, and the Transitivity of *Better Than*', *Noûs*, vol. 52, no. 4, pp. 874-899.

Nebel, J 2020, 'Rank-Weighted Utilitarianism and the Veil of Ignorance', *Ethics*, vol. 131, no. 1.

von Neumann, J & Morgenstern, O 1953, *The Theory of Games and Economic Behavior*, 3rd edn, Princeton University Press.

Nolan, D 1997, 'Impossible Worlds: A Modest Approach', *Notre Dame Journal of Formal Logic*, vol. 38, no. 4, pp. 535-572.

Nolan, D 2017, 'Causal Counterfactuals and Impossible Worlds', in H Beebee, C Hitchcock and H Price (eds), *Making a Difference*, Oxford University Press, Oxford, pp. 14-32.

Nozick, R 1969, 'Newcomb's Problem and Two Principles of Choice', in N Rescher (ed.), *Essays in Honor of Carl G. Hempel*, Reidel, Dordrecht.

Parfit, D 1984, *Reasons and Persons*, Clarendon Press: Oxford.

Peterson, M 2013, *The Dimensions of Consequentialism: Ethics, Equality, and Risk*, Cambridge University Press, Cambridge.

Peterson, M 2019, 'The St. Petersburg Paradox', in *The Stanford Encyclopedia of Philosophy*. Available at: <https://plato.stanford.edu/entries/paradox-stpetersburg/#IgnoSmalProb>

Pettigrew, R 2015a, 'Accuracy and the Credence-Belief Connection', *Philosophers' Imprint*, vol. 15, no. 16, pp. 1-20.

Pettigrew, R 2015b, 'Risk, Rationality and Expected Utility', *Canadian Journal of Philosophy*, vol. 45, no. 5, pp. 798-826.

Pettit, P 1991, 'Decision Theory and Folk Psychology', in M Bacharach and S Hurley (eds), *Essays in the Foundations of Decision Theory*, Blackwell, pp. 147-175.

Price, H 1986, 'Against Causal Decision Theory', *Synthese*, vol. 67, no. 2, pp. 195-212.

Price, H 2012, 'Causation, Chance and the Rational Significance of Supernatural Evidence', *Philosophical Review*, vol. 121, no. 4, pp. 483-538.

Quiggin, J 1982, 'A Theory of Anticipated Utility', *Journal of Economic Behaviour and Organisation*, vol. 3, no. 4, pp. 323-343.

Quiggin, J 1990, 'Stochastic Dominance in Regret Theory', *Review of Economic Studies*, vol. 57, no. 3, pp. 503-511.

Quiggin, J 1994, 'Regret Theory with General Choice Sets', *Journal of Risk and Uncertainty*, vol. 8, no. 2, pp. 153-165.

Quinn, WS 1990, 'The Puzzle of the Self-Torturer', *Philosophical Studies*, vol. 59, no. 1, pp. 79-90.

Rabinowicz, W 1985, 'Ratificationism Without Ratification: Jeffrey Meets Savage', *Theory and Decision*, vol. 19, no. 2, pp. 171-200.

Rabinowicz, W 1995, 'To Have One's Cake and Eat It, Too: Sequential-Choice and Expected-Utility Violations', *The Journal of Philosophy*, vol. 92, no. 11, pp. 586-620.

Rabinowicz, 2000, 'Money Pump with Foresight', in J Almeida (ed.), *Imperceptible Harms and Benefits*, Kluwer, Dordrecht.

Rabinowicz, W 2002, 'Does Practical Deliberation Crowd Out Self-Prediction?', *Erkenntnis*, vol. 57, no. 1, pp. 91-122.

Rabinowicz, W 2009, 'Letters from Long Ago: On Causal Decision Theory and Centred Chances', in J Lars Göran, J Österberg and R Sliwinski (eds), *Logic, Ethics and All That Jazz: Essays in Honor of Jordan Howard Sobel*, Department of Philosophy, Uppsala University, Sweden.

Rabinowicz, W 2014, 'Safeguards of a Disunified Mind', *Inquiry*, vol. 57, no. 3, pp. 356-383.

Rabinowicz, W forthcoming, 'Incommensurability meets Risk' in H Anderssen and A Herlitz (eds), *Value Incommensurability: Ethics, Risk, and Decision-Making*, Routledge.

Regenwetter, M, Dana, J & Davis-Stober, CP 2010, 'Testing Transitivity of Preferences on Two-Alternative Forced Choice Data', *Frontiers in Psychology*, vol. 1, no. 148, pp. 1-15.

Regenwetter, M, Dana, J & Davis-Stober, CP 2011, 'Transitivity of Preferences', *Psychological Review*, vol. 118, no. 1, pp. 42-56.

Richter, R 1984, 'Rationality Revisited', *Australasian Journal of Philosophy*, vol. 62, no. 4, pp. 392-403.

Rinard, S 2015, 'A Decision Theory for Imprecise Probabilities', *Philosophers' Imprint*, vol. 15.

Rothschild, M & Stiglitz, J 1970, 'Increasing Risk: I. A Definition', *Journal of Economic Theory*, vol. 2, no. 3, pp. 225-243.

Rubinstein, A 1988, 'Similarity and Decision-Making Under Risk (Is There a Utility Theory Resolution to the Allais Paradox?', *Journal of Economic Theory*, vol. 46, no. 1, pp. 145-153.

Sandgren, A & Williamson, TL 2021, 'Determinism, Counterfactuals, and Decision', *Australasian Journal of Philosophy*, vol. 99, no. 2, pp. 286-302.

Savage, LJ 1972, *The Foundations of Statistics,* 2nd edn, Dover Publications, New York.

Scarsini, M 1988, 'Dominance Conditions for Multivariate Utility Functions', *Management Science*, vol. 34, no. 4, pp. 431-554.

Schick, F 1986, 'Dutch Bookies and Money Pumps', *The Journal of Philosophy*, vol. 83, no. 2, pp. 112-119.

Schoenfield, M 2014, 'Decision Making in the Face of Parity', *Philosophical Perspectives*, vol. 28, no. 1, pp. 263-277.

Sen, A 1970, *Collective Choice and Social Welfare*, Holden-Day, San Francisco.

Sen, A 1971, 'Choice Functions and Revealed Preference', *The Review of Economic Studies*, vol. 38, no. 3, pp. 307-317.

Seidenfeld, T 1984, 'Comments on Causal Decision Theory' *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol. 2, pp. 201-212.

Seidenfeld, T 1994, 'When Normal and Extensive Form Decisions Differ', in D Prawitz, B Skyrms & D Westerstål (eds), *Logic, Methodology and Philosophy of Science IX*, Elsevier, North Holland, pp. 451-463.

Shafir, E & Tversky, A 1992, 'Thinking Through Uncertainty: Nonconsequential Reasoning and Choice', *Cognitive Psychology*, vol. 24, no. 4, pp. 449-474.

Sen, A 1971, 'Choice Functions and Revealed Preference', *Review of Economic Studies*, vol. 38, no. 3, pp. 307-317.

Sen, A 1997, *Choice, Welfare and Measurement*, Harvard University Press, Cambridge, Massachusetts.

Skiadas, C 1997a, 'Conditioning and Aggregation of Preferences', *Econometrica*, vol. 65, no. 2, pp. 347-367.

Skiadas, C 1997b, 'Subjective Probability under Additive Aggregation of Conditional Preferences', *Journal of Economic Theory*, vol. 76, no. 2, pp. 242-271.

Skyrms, B 1982, 'Causal Decision Theory', *The Journal of Philosophy*, vol. 79, no. 11, pp. 695-711.

Skyrms, B 1984, *Pragmatics and Empiricism*, Yale University Press, New Haven.

Slovic, P & Peters, E 2006, 'Risk Perception and Affect', *Current Directions in Psychological Science*, vol. 15, no. 6, pp. 322-325.

Slovic, P, Peters, E, Finucane, M & MacGregor, D 2002, 'Affect, Risk, and Decision-Making', *Health Psychology*, vol. 24, no. S4, pp. 35-40.

Smith, NJJ 2014, 'Infinite Decisions and Rationally Negligible Probabilities', *Mind*, vol. 125, no. 500, pp. 1199-1212.

Snedegar, J 2017, *Contrastive Reasons*, Oxford University Press, Oxford.

Sobel, J 1986, 'World Bayesianism: Comments on the Hammond/McClennen Debate', in B Munier (ed.), *Risk, Decision and Rationality*, D. Reidel, Holland, pp. 537-542.

Sobel, J 1988a, 'Utility Theory and the Bayesian Paradigm', in *Taking Chances: Essays on Rational Choice*, Cambridge University Press, Cambridge.

Sobel, J 1988b, 'Useful Intentions', in *Taking Chances: Essays on Rational Choice*, Cambridge University Press, Cambridge.

Sobel, J 1988c, 'Expected Utilities and Rational Actions and Choices', in *Taking Chances: Essays on Rational Choice*, Cambridge University Press, Cambridge.

Sobel, J 1988d, 'Infallible Predictors', *Philosophical Review*, vol. 97, no. 1, pp. 3-24.

Spencer, J & Wells, I 2019, 'Why Take Both Boxes?', *Philosophy and Phenomenological Research*, vol. 99, no. 1, pp. 27-48.

Spohn, W 1977, 'Where Luce and Krantz Do Really Generalize Savage's Decision Model', *Erkenntnis*, vol. 11, pp. 113-134.

Steele, K 2010, 'What are the Minimal Requirements of Rational Choice? Arguments from the Sequential-Decision Setting', *Theory and Decision*, vol. 68, pp. 463-487.

Stern, R 2017, 'Interventionist Decision Theory', *Synthese*, vol. 194, no. 1, pp. 4133-4153.

Stefánsson, HO & Bradley, R 2019, 'What is Risk Aversion', *British Journal for the Philosophy of Science*, vol. 70, no. 1, pp. 77-102.

Tarsney, C 2020, 'Exceeding Expectations: Stochastic Dominance as a General Decision Theory', *Global Priorities Institute Working Papers*, no. 3.

Temkin, L 2012, *Rethinking the Good*, Oxford University Press, Oxford.

Thoma, J 2019, 'Risk Aversion and the Long Run', *Ethics*, vol. 129, no. 2, pp. 230-253.

Thoma, J 2021, 'Judgementalism about Normative Decision Theory', *Synthese*, vol. 198, pp. 6767-6787.

Tversky, A 1969, 'Intransitivity of Preferences', *Psychological Review*, vol. 76, no. 1, pp. 31-48.

Van Fraasen, B 1984, 'Belief and the Will', *Journal of Philosophy*, vol. 81, no. 5, pp. 235-256.

Voorhoeve, A & Binmore, K 2006, 'Transitivity, the Sorites Paradox, and Similarity-Based Decision-Making', *Erkenntnis*, vol. 64, no. 1, pp. 101-114.

Weatherson, B 2012, 'Knowledge, Bets, and Interests', in J Brown and M Gerken (eds), *Knowledge Ascriptions*, Oxford University Press, Oxford, pp. 75-103.

Wedgewood, R 2013, 'Gandalf's Solution to Newcomb's Problem', *Synthese*, vol. 190, no. 4, pp. 2643-2675.

Wells, I 2019, 'Equal Opportunity and Newcomb's Problem', *Mind*, vol. 128, no. 510, pp. 429-457.

Weirich, P 1983, 'A Decision Maker's Options', *Philosophical Studies*, vol. 44, no. 2, pp. 175-186.

Weirich, P 1984, 'The St. Petersburg Gamble and Risk', *Theory and Decision*, vol. 17, no. 2, pp. 193-202.

Weirich, P 1985, 'Decision Instability', *Australasian Journal of Philosophy*, vol. 63, no. 4, pp. 465-472.

Weirich, P 2020, 'Risk as a Consequence', *Topoi*, vol. 39, pp. 293-303.

Wilkinson, H Forthcoming, 'In Defence of Fanaticism', *Ethics*.

Williams, R 2014, 'Decision-Making Under Indeterminacy', *Philosophers Imprint*, vol. 14.

Williamson, TL 2021, 'Causal Decision Theory is Safe from Psychopaths', *Erkenntnis*, vol. 86, no. 3, pp. 665-685.

Williamson, TL & Sandgren, A forthcoming, 'Law-Abiding Causal Decision Theory', *The British Journal for the Philosophy of Science*.

Zynda, L 2000, 'Representation Theorems and Realism about Degrees of Belief', *Philosophy of Science*, vol. 67, no. 1, pp. 45-69.