Choice and Credence in Context

by

Calum McNamara

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2024

Doctoral Committee:

Professor James M. Joyce, Chair
Professor Gordon Belot
Professor Ezra Keshet
Professor Sarah Moss
Professor Brian Weatherson

Calum McNamara

camcnam@umich.edu

ORCID iD: 0000-0003-1032-073X

For Maria, who made it possible.

Acknowledgments

As the saying goes, it takes a village. This dissertation wouldn't have been written if it weren't for the support, guidance, and encouragement of many different people. Thanks to all of you.

My greatest intellectual debt is owed to my advisor, Jim Joyce. Even before I arrived at Michigan, I was a fan of Jim's work (I wrote my undergraduate thesis on accuracy-first epistemology). And after six years in Michigan's department, my admiration for Jim hasn't diminished by one bit. Throughout, he's been a wonderful advisor: Jim has both given me the freedom to explore ideas on my own, and he's steered me back on track whenever I looked in danger of getting lost. His influence on this dissertation is hard to overstate. I doubt Jim will agree with everything I say here. But I know he agrees with some of it. And this is about the highest compliment that I could've received on my work. Jim: thanks for everything.

I also owe a huge debt of gratitude to the other members of my dissertation committee: Gordon Belot, Sarah Moss, Brian Weatherson, and Ezra Keshet.

Gordon is the ultimate philosophical gadfly. At various points during grad school, he's challenged my—mostly Bayesian—presuppositions in helpful and constructive ways. Gordon has also given me what are perhaps the best comments on papers that I could've asked for—not to mention correcting many of my fallacious proofs. This dissertation has been immeasurably improved by his input (something that Gordon might be surprised to hear, given his reservations about conditionals). More than this, however, Gordon has always served to remind me of something crucial about philosophy: whatever else it is, it's a lot of fun.

Sarah took me under her wing more-or-less straight away when I arrived at Michigan. Over the last six years, she's served as a model for how to approach philosophy, both personally and professionally. Sarah also gave me a very important piece of advice early on: "When it comes to doing philosophy, just try to say things that are true". I don't think I've gotten to the bottom of the issues I discuss in this dissertation. But I have tried to say only things that are true. And approaching things in this way is something that I owe, very much, to Sarah.

Brian, despite his immense intelligence and truly impressive knowledge of the literature, has the enviable gift of making a person feel completely comfortable floating ideas in conversation. If it weren't for Brian's encouragement at various points during grad school, I probably wouldn't have had the guts to explore several of the ideas I discuss in these pages. What's more, Brian has taught me basically everything I know about a range of non-philosophy subjects: game theory, the French revolution, World War II fighter planes, and US immigration law—he really is a modern day renaissance man. And his intellectual curiosity is something I aspire to.

Finally, Ezra laid the foundations for my interest in formal semantics way back in 2019, encouraging me to do work in this area. Since then, he's indulged me by running reading groups on topics that I requested—notably, a Summer semantics reading group on epistemic modals and conditionals. More-or-less everything I know about formal semantics is due to Ezra. And I hope he'll recognize his influence at various points in this dissertation.

I also need to thank the other members of Michigan's faculty, for help and support over the years.

In particular, I'd like to single out: Dave Baker, Tilman Börgers (in economics), Victor Caston, Maegan Fairchild, Allan Gibbard, Renee Jorgensen, Ishani Maitra, David Manley, and Peter Railton (who introduced me to the Desire-as-Belief thesis, discussed in Chapter 3); Laura Ruetsche, Tad Schmaltz, and Gabriel Schapiro (even if we never did cross paths at Michigan!); Janum Sethi, Eric Swanson (especially for his tireless work as placement director); and Rich Thomason (for interesting discussions about conditionals).

The grad students at Michigan are also deserving of a great deal of my gratitude. In particular, there's: Abdul Ansari, Mitch Barrington, Kevin Blackwell, Jason Byas, and Francisco Calderón (the lattermost of whom has been one of my closest friends in grad school, despite our wildly divergent philosophical views); Paul de Font-Reaulx, Gillian Gray, Malte Hendrickx, and Josh Hunt (a dear friend, despite his one-boxing tendencies); Ina Jängten (an honorary Michigander); Gabrielle Kerbel (my philosophical little sister); Lorenzo Manuali, Eduardo Martinez (especially for putting up with me as a co-author); Cameron McCulloch (my fellow Scotsman); Ariana Peruzzi, Josh Petersen, Laura Soter, Angela Sun, and Brett Thompson (Brett, sorry for sending you so many papers); Sarah Valdman, Alison Weinberger, Margot Witte, and Elise Woodard (especially for all her help early on); Sophia Wushanley, and Glenn Zhou. (Some of these people now have flourishing academic careers of their own. But in my heart, they'll always be fellow Wolverines.)

Thanks also to Michigan's immensely helpful staff: Mia Arnold, Judith Beck, Kelly Campbell, and Shelley Anzalone. And thanks, especially, to Carson Maynard for always coming to the rescue as Grad Program Coordinator. (Seriously: I wouldn't be finishing my PhD now if it weren't for Carson.)

Last but not least, thanks to all of the undergraduates who I've had the pleasure of teaching while at Michigan. I think undergrads often don't realize the effect they have on their teachers. But in a few cases, I've found myself learning almost as much from my students as I've been able to teach to them. Better yet, I'm lucky to now call a few of my former students friends. (Joyce fan club: I'm looking at you.)

Outside of Michigan, I've benefitted from interactions with grad students at other departments. In particular, I'd like to thank all of the grad students at NYU and Princeton, where I visited during the Fall Semester of 2022; and the grad students at UC Irvine, with whom I interacted while taking part in the epic Formal Epistemology Reading Group, which ran throughout the whole of the pandemic. Special thanks go to Cristina Ballarini, Chris Bottomley, Clara Bradley, Sam Cantor, and Simon Dietz (who's been my philosophical better even while we were undergrads); Hüseyin Güngör (who was one half of my support network while on the job market); Daniel Herrmann (especially for early conversations about Ahmed's "deterministic cases"); Saira Khan, and Bar Luzon (who was the other half of my support network while on the market); Aydin Mosheni, Adrian Ommunsen, Aidan Penn, Richard Roth, Gerard Rothfus, Ethan Russo, and Patrick Wu.

I'm also lucky to have been mentored by some amazing philosophers at other departments. Mostly obviously, I'd like to thank all of the philosophers at King's College London—where I was an undergraduate—for their early support and encouragement, and especially Julien Dutant, Clayton Littlejohn, and Mark Textor, each of whom played a crucial role in getting me into grad school. There are also the following people, all of whom provided me with crucial feedback or encouragement at some point(s) over the last six years: Arif Ahmed, Boris Babic, Andrew Bacon, Zach Barnett, David Boylan, Catrin Campbell-Moore, Jessica Collins, Guillermo Del Pinal ("Sparta for all"), Cian Dorr (especially for mentoring me while I was a visitor at NYU), Adam Elga (for encouraging me to explore the contextualist approach to CDT that I discuss in Chapter 2), Branden Fitelson, Bas van Fraassen, Dmitri Gallow (my philosophical big brother), Verónica Gómez Sánchez, Jeremy Goodman, Dan Greco, Wes Holliday (for helpful comments on co-authored work with Snow Zhang), Simon Huttegger, Harvey Lederman (for letting me ride Snow's coat-tails to a talk at UT Austin), Zoe Johnson King (for helpful advice *qua* international student when I first arrived at Michigan), Stefan Kaufman, Mikayla Kelley (especially for getting me to do karaoke), Justin Khoo, Boris Kment, Jason Konek, Alex Meehan, Eleanore Neufeld, Alejandro Perez Carballo, Richard Pettigrew, Ezra Rubinstein, Paolo Santorio, Ginger Schultheis, and, lastly, Brian Skyrms.

Extra special thanks go to Richard Bradley, Melissa Fusco, Al Hájek, Matt Mandelkern, and Snow

Zhang. Richard read nearly all of this dissertation while it was still in its early stages, and provided me with a crucial letter of recommendation for the job market—something I'm sure made all the difference to my landing some amazing jobs. Similarly, Melissa provided me with feedback while this project was still in its infancy, and has been an immensely helpful interlocutor as it's evolved. Al has been a champion of my work for at least a year now. And he was kind enough to offer my a post-doc at ANU (which, even now, I'm sad to say I had to decline). Matt, quite literally, got me interested in the sequence semantics for conditionals in the first place. And over the last two years, he's played an immensely helpful role in shaping my ideas. Also, Matt is one of the most responsive emailers I've ever interacted with—something that's made all the more impressive by his immense publication record. (Seriously, Matt, how do you find the time?) Finally, Snow's influence on me—both personally and philosophically—is hard to over-estimate. There are so many places in this dissertation where she's had an impact on my ideas—indeed, Chapter 4 had its genesis in early conversations with Snow. But aside from this, Snow is one of the kindest, and most generous, people I've ever interacted with, and I feel lucky to call her a friend (even if we never can seem to finish our joint paper!).

Thanks to an anonymous referee at *Philosophy and Phenomenological Research*, who provided very helpful feedback on Chapter 2 of this dissertation; thanks to the philosophers at Yale and Indiana, who I'm excited to call colleagues soon; and thanks to audiences at Berkeley, Chicago, Michigan, UConn, UMass, UT Austin, and various APAs for their feedback on this—or related—material.

Thank to my family, both immediate and extended, for their unwavering support: Mum; Dad; Rhian, Sandy, and Charlie; George and Mela; Anna and Matt (and kids); Fiona ("Ham and potatoes supper"); Felix, Emma, and children. And thanks also to my non-philosophy friends in Ann Arbor.

No thanks go to Mishka (my cat), who would rather meow for my attention than let me get any work done. (Nevertheless, she is very cute.)

And lastly, and most importantly: thanks go to Maria. When I first met Maria, I was in a transitory phase of my life, having recently given up on the idea of a career in classical music. At that time, I wasn't sure what I would do next. And Maria would've been perfectly justified in thinking that I wouldn't amount to very much. Still, she supported me and encouraged me—even when my interests turned to a field with even worse job prospects than classical music. Also, when I was admitted to grad school, Maria gave up a good job in London so that I could pursue my dream of doing philosophy. In the truest possible sense, then, this dissertation wouldn't have been written if it weren't for her. And for that reason, Maria, I'm dedicating it to you.

Contents

List of Figures

List of Tables

Abstract

This dissertation is about the role that conditionals play in uncertain reasoning and deliberation. Specifically, I attempt to show that, by appealing to a particular *semantics* for conditionals—a contexutalist, *sequence semantics*, which has recently become popular in philosophy of language—several open problems in decision theory and epistemology can be solved.

Chapter 1 is introductory. I set out the semantic view of conditionals in question, and I describe some of its historical background.

Chapter 2 turns to a striking problem faced by causal decision theorists. A popular formulation of causal decision theory (CDT) appeals to counterfactual conditionals. However, the standard theory of these conditionals has unintuitive consequences in deterministic worlds. In particular, it says that if anything—including the choice you make—were different in the present, then either the laws or nature would be violated, or the distant past would be changed. And as several authors have recently shown, it's easy to transform this consequence of the standard theory of counterfactuals into full-blown counterexamples to CDT. In response to these counterexamples, I develop a contextualist version of CDT, which makes use of the sequence semantics. I then show that the deterministic counterexamples don't arise for my version of CDT.

In Chapter 3, I deal with a different puzzle, about whether or not the so-called *Desire-as-Belief (DAB) thesis* is consistent with decision theory—something that famous arguments of David Lewis seem to show isn't the case. Once again, I show that, if we understand the DAB thesis in a contextualist way—and spell it out using the sequence semantics—then Lewis's arguments against that thesis don't go through. In fact, we can prove a *tenability result* for the DAB thesis, which shows that it's compatible with decision theory after all.

Finally, in Chapter 4, I transition from decision-theoretic issues to epistemological ones. More precisely, I tackle the question of how our credences should change when we learn indicative conditionals. Several famous cases in the literature—notably, Bas van Fraassen's *Judy Benjamin* problem—seem to show that the standard Bayesian update rules deliver implausible results when we learn conditionals of this kind. However, in the chapter, I show that, if we adopt the sequence semantics, then the Bayesian update rules turn out to deliver the correct results after all. Better still, alternatives to these rules which have been put forward in the literature turn out to be equivalent to the Bayesian rules in my framework—at least in many contexts. Thus, what we end up with is a nice, unified account of how rational agents should update on conditional information: one which fits in well with recent work on the semantics of conditionals. My proposal also relates, in interesting ways, to discussions that have been happening elsewhere in the literature, like discussions about the tenability of the notorious *Stalnaker's thesis*.

Chapter 1

A Note on the Semantics of Conditionals

1.1 Introduction

We appeal to conditionals all the time, not least in uncertain reasoning and deliberation. For example, consider these sentences:

(1) If I leave my umbrella at home, then I might get wet on my way to work.

(2) If the Butler didn't do it, then the Gardener probably did.[1]

(3) If Oswald hadn't shot Kennedy, then I'm pretty sure no one else would've done so.[2]

(4) If I take both boxes, then, no matter what's in the opaque box, I'll get $\$1,000$ more than I would if I took only that box.[3]

All of these sentences seem like perfectly natural things to say to yourself in situations involving uncertainty—be they deliberational or epistemic. And I bet you can come up with additional examples of your own, to supplement mine.

Still, despite how central conditionals are to our epistemic and practical lives, they remain something of a mystery. For example, there's widespread disagreement in philosophy (and linguistics) about the *truth-conditions* for conditionals, with some authors denying that they have truth-conditions in the first place.[4] Additionally, among those who think that conditionals really do have truth-conditions, there's disagreement about whether those truth-conditions are the same for indicatives and subjunctives (viz., counterfactuals), or whether they're different.[5] And there's disagreement about the role that *context* plays

[1] Cf. Stalnaker (1975), Edgington (1995).

[2] Cf. Adams (1970).

[3] Cf. Nozick (1969) and chapter 2 below.

[4] See, for example, Adams (1975), Gibbard (1981), Edgington (1995), Bennett (2003), Moss (2015, 2018), or Ciardelli and Ommundsen (2022), all of whom deny that indicative conditionals have truth-conditions. Additionally, Edgington and Skyrms (1980b) deny that counterfactuals have truth-conditions. See the next footnote for a brief illustration of the difference between indicative conditionals and counterfactuals.

[5] In what follows, I'm going to use the terms 'subjunctive conditional' and 'counterfactual' interchangeably. I'll also largely take for granted that you, the reader, are familiar with the distinction between indicative conditionals and subjunctives. Just to be sure, however, the following sentence is an indicative conditional, whereas (3) in the main text is a paradigmatic example of a subjunctive (counterfactual) conditional:

(i) If Oswald didn't shoot Kennedy, someone else did.

Hopefully, these examples are enough to give you an intuitive grasp of the distinction between indicatives and subjunctives. Note

in informing these truth-conditions, too.

This dissertation starts by assuming a particular semantic view about conditionals, and then uses that view to address some open problems in decision theory and epistemology. The view in question is one inspired by the work of Stalnaker (1968, 1970, 1975, 1981a, 1984) and van Fraassen (1976), and developed more recently by Bacon (2015), Khoo and Santorio (2018), Khoo (2022), Schultheis (2023), and Mandelkern (forthcoming), among others.[6] In the three chapters that follow this one, I'll largely be taking this semantic view as a given. But, in this introductory chapter, I want to take a bit of time to explain, motivate, and defend that view. What I'll say here isn't meant to be decisive—my remarks won't constitute anything like knock-down objections to other semantic views, for example. But hopefully, they'll help to situate the Stalnaker-van-Fraasen-inspired view in a wider context, and show why it's plausible and useful. Besides, the fact that adopting this view can be used to solve long-standing problems in decision theory and epistemology functions, I think, as an indirect argument in its favor. One of the things we do in philosophy (and science), after all, is judge a theory by its fruits.

In §1.2, I'll give a short overview of Stalnaker's original semantic theory of conditionals (1968), which in some sense lays the groundwork for the semantic theory I'll eventually adopt. In §1.3, I'll discuss some of the advantages of Stalnaker's view, especially with how it relates to probability. In §1.4 I'll introduce Lewis's famous *triviality results* and their spin-offs, and say what we should think about the Stalnakerian view in light of these results. Finally, in §1.5 I'll say how the view I adopt in subsequent chapters of the dissertation can be used to get around the triviality results, and why this fact makes for a strong point in its favor. I close the chapter in §1.6 with some prefatory remarks about directions for future research.

1.2 Stalnaker's Semantics

Frank Ramsey had many good ideas,[7] but a particularly good, and simple, one is the following:

> If two people are arguing 'if A will C?' and are both in doubt as to A, they are adding A hypothetically to their stock of knowledge and arguing on that basis about C... We can say that they are fixing their degrees of belief in C given A. (Ramsey, 1929, p. 155, with trivial changes of notation)

The idea contained in this paragraph—which has come to be known as the *Ramsey test* for conditionals—is that a conditional is "believable" or "acceptable" to you just in case its consequent seems likely *on the supposition of its antecedent.* At a first pass, that seems correct. After all, consider this sentence:

(5) If I flip this fair coin, it will land heads.

To what degree are you willing to believe or "accept" this conditional?[8] Presumably, your answer is something like '50%'. And that, plausibly, is just the degree to which you believe that the coin will land heads,

also that—despite their name—I won't be assuming that counterfactuals have false antecedents in what follows.

[6] Important additional contributions to the view of conditionals I'll develop have been made by McGee (1989), Stalnaker and Jeffrey (1994), Kaufmann (2004, 2005, 2015), Bradley (2012, 2017), and Goldstein and Santorio (2021). The authors listed in the main text develop a semantic view of conditionals that's closest in spirit to the one I'll make use of here—mostly because they're explicitly *contextualists*—whereas the authors listed in this footnote are less forthcoming about how their view fits with contextualism (and, indeed, some of them are explicitly anti-contextualist).

[7] See Misak (2020) for a recent biography or Ramsey, which makes clear just how many good ideas Ramsey had in his short life.

[8] Two things. First, I'm not going to say here exactly what "acceptance" amounts to. But in §1.3 we'll see two different precisifications of the Ramsey test, which plausibly capture the notion of "acceptance". Second, to keep things simple, I'll sometimes speak sloppily in the main text, saying things like 'To what degree do you believe this *sentence*?' Really, it's the *contents* of sentences to which we attach credences (degrees of belief). But it'll simplify the discussion if I'm allowed to speak in the sloppy way.

under the supposition that it's flipped.

Later on, we'll encounter two ways of making the Ramsey test precise (§1.3). But before that, let's see how the Ramsey test inspired Stalnaker's (1968) semantics for conditionals.[9]

The Ramsey test describes a relationship between the "acceptance-" or "belief-conditions" for conditionals, on the one hand, and our willingness to accept or believe the consequent of such a conditional under the supposition of the antecedent, on the other. As Stalnaker points out, however, if we want to turn this idea into a *semantic* theory of conditionals, then we need a way of connecting Ramsey's claims about states of acceptance or belief and suppositions to *truth-conditions*. As Stalnaker puts it:

> [The Ramsey test answers] the question, 'How do we decide whether or not to believe a conditional statement?' [But] the problem is to make the transition from belief conditions to truth conditions; that is, to find a set of truth conditions for statements having conditional form which explains why we use the method we do use to evaluate them. (p. 33)

How, then, should we make the transition that Stalnaker has in mind?

Here's how he says we should proceed:

> The concept of a *possible world* is just what we need to make this transition, since a possible world is the ontological analogue of a stock of hypothetical beliefs. The following set of truth conditions, using this notion, is a first approximation to the account that I shall propose: Consider a possible world in which A is true, and which otherwise differs minimally from the actual world. 'If A, then C' is true (false) just in case C is true (false) in that world. (pp. 33-34, with a trivial change of notation)

In slogan form, then, Stalnaker's view is that a conditional 'If A, then C' is true just in case the "minimally different" world in which the antecedent is true is one in which the consequent is true.

Of course, as Stalnaker admits, this is only a rough approximation of his final semantic theory of conditionals. (Note also that this rough idea is supposed to apply to both indicative conditionals and counterfactuals, making Stalnaker's semantics is a "uniform" semantics. I'll say more about how we distinguish between these conditionals in Stalnaker's semantics later on.) Thus, what we need to do now is see how that rough approximation can be made more precise.

To start with, then, let's suppose that we have a set, $\mathcal{W}$, of possible worlds, each member of which is the "ontological analogue of a stock of hypothetical beliefs". For simplicity, I'll assume throughout that $\mathcal{W}$ is finite (and this goes for the dissertation as a whole). We can then think of *propositions* as subsets of $\mathcal{W}$. So, the set of all propositions is the power set of $\mathcal{W}$, which I'll write '$\mathcal{P}(\mathcal{W})$'. A proposition, A, is *true* at a possible world $w \in \mathcal{W}$ just in case $w \in A$. (In §1.5, I'll revise the foregoing definition of 'proposition'. But for now, it's the one I'll stick with.)

The next thing we need to do is introduce the notion of a *selection function*. Formally, this is a function $f : \mathcal{P}(\mathcal{W}) \times \mathcal{W} \to \mathcal{W}$ which takes a proposition A and a world w as its arguments, and maps these to a possible world, $f(A, w)$, which corresponds to what Stalnaker calls the "minimally different" A-world to w. Stalnaker himself doesn't say much about this function, other than that it obeys some abstract constraints. In particular, he thinks that selection functions should satisfy the following conditions:[10]

(i) **Success**. $f(A, w) \in A$.

(ii) **Strong Centering**. If $w \in A$, then $f(A, w) = w$.

[9] Stalnaker's theory is presented more formally in Stalnaker and Thomason (1970). A similar semantics for conditionals—or at least, for *subjunctive* conditionals—is given by Lewis (1973b). As we'll see, Stalnaker's theory differs from Lewis's in a few key aspects. And the ways in which it differs also function as the reasons that I prefer it.

[10] Stalnaker also has a condition, Absurdity, which tells us what the selection function outputs when A is the empty set. This particular constraint won't be important here. So I'll ignore it.

(iii) **Reciprocity**.[11] If $f(A, w) \in B$ and $f(B, w) \in A$, then $f(A, w) = f(B, w)$.

Roughly speaking, then, Success says that the minimally different A-world to w should *be* an A-world. Strong Centering says that, if $w \in A$, then w is the minimally different world to itself. Reciprocity is needed to validate a host of compelling inference patterns involving conditionals. And a bit of reflection shows that all of these conditions are required if $f(A, w)$ is going to correspond, in any intuitive sense, to our pre-theoretic notions of "minimal difference". I'll assume that selection functions satisfy all of these constraints, in all of what follows.

On top of the constraints (i)–(iii), Stalnaker (1975) says that selection functions should satisfy a further constraint, at least when the conditionals in question are *indicative* conditionals:

(iv) **Indicative Constraint**. Let $E \subseteq \mathcal{W}$ be the set of your epistemically possible worlds. Then, if $E \cap A \neq \varnothing$, it follows that $f(A, w) \in E$.

The idea underpinning this constraint is that, in the case of indicative conditionals (but not in the case of counterfactuals), the minimally different A-world to an epistemically possible world should itself be epistemically possible. This constraint helps to capture some of the idea that indicative conditionals are about *epistemic* possibilities, whereas subjunctive conditionals are about *causal*, or *metaphysical*, possibilities (see Chapter 2, as well as, e.g., Stalnaker (1975), Mandelkern (2018, forthcoming), Khoo (2022), or Schultheis (2023) for further discussion, as well as for refinements of this idea).

Now, aside from the constraints (i)–(iv), Stalnaker doesn't say much about what it means for a world to count as the minimally different A-world to a given world w. Instead, he merely says that the constraints (i)–(iv) "on the selection function are necessary in order that this account be recognizable as an explication of the conditional, but they are of course far from sufficient to determine the function uniquely" (p. 36). In subsequent work, he fleshes out this point out by appealing to the notion of *context*. In particular, he says that it's largely a matter of context which selection function(s) is (are) acceptable: "Everyone agrees that conditional statements are context-dependent" and thus "the relevant… selection function for their interpretation may depend on the subject matter of the antecedent and consequent, on features of the local conversational context such as the presumed interests of the participants of the conversation", etc. (MS, pp. 6-7). I myself have some ideas about how context works to picks out "admissible" selection functions, at least in the case of counterfactuals. And I discuss these ideas in detail in Chapter 2.

For now, however, let me just state Stalnaker's semantics precisely, trusting that we have an intuitive grip on how context might fill in the details on the notion of "minimal difference". Given a selection function, f, the semantics is:

Stalnaker Semantics. $[\![\text{If } A, \text{ then } B]\!]^w = 1$ if and only if $f(A, w) \in B$.

Roughly speaking, then, Stalnaker's semantics says that the sentence 'If A, then B' is true at the world w just in case the selected A-world is a B-world. (Strictly speaking, I should add a context variable, c, to the statement of this semantics. But I'll ignore that here, for simplicity.)

Now, this semantics makes very plausible empirical predictions about the truth-values of conditionals, as many authors have noted (e.g., Lewis, 1973b, 1979; Moss, 2012; Mandelkern, forthcoming). To illustrate this anyway, however, consider a famous example from Jackson (1977):

(6) If Fred had jumped off the roof of the Empire State Building, he would've died.

In this case, it seems overwhelmingly plausible that the sentence (6) is true. And that seems vindicated by a natural application of Stalnaker Semantics. After all, assuming that Jones didn't actually jump off

[11] In the literature, this constraint is sometimes called 'CSO'. But no one has been able to tell me what that means. I take the name 'Reciprocity' from Mandelkern (forthcoming).

the roof—so the antecedent of (6) is false—the minimally different world at which he *did* jump off the roof seems like it should be one in which there's no net in place to catch him, where gravity works the same as we're used to, and so on. Thus, at this world, it seems overwhelmingly likely that Jones would die, if he jumped. So, according to Stalnaker Semantics, the sentence (6) comes out true.[12]

On top of its plausible empirical predictions, Stalnaker's semantics gives rise to an extremely plausible *logic* for conditionals. To take a well-known example to illustrate this, consider the following inference pattern:[13]

P If Fred had got his son a puppy for Christmas, the son would've been delighted.

C If Fred had got his son a puppy for Christmas, and then strangled it, the son would've been delighted.

Clearly, the inference from P to C is invalid. But surprisingly, many well-known theories of (natural language) conditionals turn out validate it. One example of this is the *strict conditional theory*, according to which a counterfactual $A > C$ is merely a necessitated *material* conditional, $\Box(A \supset C)$. Such a view has been defended, for example, by Von Fintel (2001), Gillies (2007), T. Williamson (2020), and Moss (MS), among others.[14] (By the way, I'm going to write '$A > C$' for the proposition expressed by a sentence 'If A, then C' or 'If A, would C' in what follows. I'll use the same operator, $>$, for both indicatives and subjunctives, since—again—Stalnaker's semantics is a uniform semantics. However, context will always make clear which kind of conditional I have in mind.)

Stalnaker's theory, however, doesn't validate the inference from P to C above, which is an instance of so-called *Antecedent Strengthening*. It's also easy to see why: according to Stalnaker's theory, P is true just in case the minimally different world at which Fred got his son a puppy is one at which the son is delighted. But *that* minimally different world needn't be the same minimally different world at which Fred bought the puppy and strangled it. So, the inference from P to C fails, according to Stalnaker. And this seems like a point in favor of his semantics.

There are many other examples like this one, which illustrate the plausibility of Stalnaker's semantics and the logic for conditionals it gives rise to. However, I won't say more about those examples here—save for one important exception. Before we move on, I want to briefly discuss an important—and notorious—point about Stalnaker's semantics. This is that it validates the controversial principle known as *Conditional Excluded Middle* (CEM). This principle is the following:[15]

Conditional Excluded Middle. $\vDash (A > C) \vee (A > \neg C)$.

In words, CEM says that sentences like the following are logical truths:

(7) If I had flipped the fair coin, it would've landed heads, or if I had flipped the fair coin, it would've landed tails.

[12] I'm skating over some details here. For reasons I can't get into in this dissertation, my considered view is that, contrary to what I've just said, the sentence (6) is actually *indeterminate* in truth-value, rather than true. I discuss my reasons for thinking this at length in McNamara (MS-c). However, the view I put forward there is controversial. And nothing I say in this dissertation depends on it.

[13] I'm taking the following example from Dorr and Hawthorne (MS). Apologies, also, for the violent nature of the preceding examples. I'm not sure why it is that examples like this one seem to be so popular in the semantics literature. Perhaps it's because they make the relevant points more forcefully than other examples.

[14] Let me say here that most strict conditional theorists agree that the inference from P to C in the main text is bad. To explain this, however, they say that this particular inference involves a subtle context shift. And it's only when context is held fixed that inferences like this are allowed. See, e.g., Von Fintel (2001) or Gillies (2007) for more on this.

[15] I'm stating this principle as a *semantic* principle, whereas Stalnaker takes it as an axiom in his logic for conditionals, C2. Nothing much turns on this.

At a first pass, (7) really does seem like a logical truth. So, prima facie, it seems like a good thing that Stalnaker's Semantics validates CEM.

In fact, however, CEM has been the subject of much debate in the philosophical literature (as well as in semantics). Most of the worries stem from its *metaphysical* implications. For example, David Lewis (1973b) famously criticized CEM in this way, saying that, if CEM is valid, then one or the other of the disjuncts in (7) is determinately true at each possible world. In other words, if CEM is valid, then for each world w, we have either that:

(8) If I had flipped the fair coin, it would've landed heads,

or else:

(9) If I had flipped the fair coin, it would've landed tails.

But how can that be, if the coin is chancy? As Alan Hájek (MS) states this worry:

> [T]o say that there is a fact of the matter of how the toss would land is to deny that the coin is a chancy system… A further fact that would steer the process one way rather than another seems wholly mysterious—and if it existed, the process would not be chancy after all, defeating the point of the example… (pp. 7-9)

Thus, according to Lewis and other critics of CEM—like Hájek—examples like this one should lead us to reject that principle.

Now, there are well-known responses to these metaphysical worries about CEM in the literature. Famously, for instance, Stalnaker (1981a) responded to Lewis by saying that, while the sentence (7) is determinately true at each possible world, that's consistent with thinking that each of its disjuncts is *indeterminate* in truth-value at each world. (Remember this point: I'll return to it in §1.5.) There are other responses to Lewis's worries, in addition to this one (see, e.g, Hawthorne, 2005).[16] But I want to defer discussion of those replies for the moment. Instead, let me now focus on a different response to the worries about CEM. This one says that, in rejecting CEM in light of its metaphysical implications, authors like Lewis (and Hájek) have got things back-to-front. To see what I mean, consider an inference pattern like the following:[17]

P No student would have passed if he goofed off.

C Every student would have failed if he goofed off.

An inference like this seems perfectly valid. But it turns out that it doesn't go through without CEM. The same is true for examples involving negations, rather than quantifiers:[18]

P It's not the case that, if Maria was in Ann Arbor, I'd be in San Francisco.

C If Maria was in Ann Arbor, I wouldn't be in San Francisco.

And there are other inutitive examples as well. In each of these cases, the relevant inferences seems perfectly valid. But they're not inferences we can make if we reject CEM. Thus, the metaphysically-influenced objections to CEM seem a bit off target. As Matthew Mandelkern puts this point:

> [Lewis's objection to CEM] is a dialectically funny objection… [What we're looking for is] an explanatory theory of conditionals. If reflective usage conflicts with that theory, that is evidence that the theory is wrong. CEM appears valid; it is up to us to find a theory of conditionals that makes sense of this. (forthcoming, p. 123)

[16] Again, I lay out some thoughts on this discussion in my MS-c.

[17] The following examples are due to Higginbotham (1986, 2003).

[18] I owe the particular example below to Paolo Santorio. For similar examples, see Mandelkern (forthcoming).

In short, then, Mandelkern's point—which I'm agreeing with—is that, when it comes to finding a plausible theory of conditionals, we should want our model theory to respect our judgments about inferences involving conditionals, and not the other way around. Stalnaker's theory respects this order of priority. The objections to CEM given by Lewis and others do not.

Now, as I said before, these remarks in favor of CEM—and Stalnaker Semantics more broadly—aren't meant to be decisive: the battle over CEM is still raging in the literature, and no side has yet emerged as the clear victor (although my sense is that proponents of CEM currently have the edge). Instead, I'm offering these remarks merely to motivate the semantic view of conditionals that I make use of in this dissertation, which is largely inspired by Stalnaker's view.

With that said, there's a further argument in favor of CEM—and thus in favor of Stalnaker Semantics—that I think is worth noting: one that's going to be especially important later on. In my view, by far the strongest reason to want a semantics that validates CEM—like Stalnaker's—lies in how this principle interacts with *probability*. It's that issue to which we now turn.

1.3 Probability

Let's go back to Ramsey's quotation:

> If two people are arguing 'if A will C?' and are both in doubt as to A, they are adding A hypothetically to their stock of knowledge and arguing on that basis about C... *We can say that they are fixing their degrees of belief in C given A.* (1929, p. 155, with trivial changes of notation and emphasis added)

Now, earlier we noted that Ramsey's remark here ties the belief- or acceptance-conditions of a conditional to the notion of a supposition. However, his remark also suggests a way of making that idea precise. To see this, consider the emphasized sentence. There, Ramsey seems to be tying the belief- or acceptance-conditions for the conditional 'If A, then C' to the notion of *conditional probability*—specifically, the conditional probability of C given that A.

This suggests a very natural precisification of the Ramsey test. To see how it works, first suppose that your *credences*—degrees of belief—at any time can be modelled by a probability function, p. Then, perhaps the most obvious way of cashing out the Ramsey test formally is the following—a thesis which has become notorious in the literature:

> **Stalnaker's Thesis**. $p(A > C) = p(C \mid A)$ (provided that $p(A) > 0$).[19]

I'm calling this thesis 'Stalnaker's Thesis' because it was first proposed by Stalnaker (1970), as a precisification of Ramsey's idea.[20] This particular precisification has been extremely influential in the literature—and it isn't hard to see why. Indeed, there seems to be something *obviously* right about Stalnaker's Thesis. After all, consider how confident your are, again, in the sentence (5):

(5) If I flip this fair coin, it will land heads.

Once more, almost everyone I ask reports their credence here to be $1/2$.[21] And—assuming they give equal credence to the propositions *Heads* and *Tails*—this is just what Stalnaker's thesis requires. So, Stalnaker's thesis seems to get things right in cases like this one.

[19] In this definition, I'm assuming the standard ratio formula for conditional probability. That is, $p(C \mid A) = p(A \wedge C)/p(A)$, assuming $p(A) > 0$. (If $p(A) = 0$, then we let the conditional probability be undefined.)

[20] Stalnaker's Thesis sometimes goes by other names in the literature. For example, some authors refer to it as *Adams' Thesis*, while others simply refer to it as *The Thesis*. For what it's worth, I think the name 'Adams' Thesis' is a misnomer, with the reason being that Adams himself merely though the *assertability* of a conditional goes by its conditional probability. On his view, conditionals don't have truth-conditions, and so the relevant kind of probability involved in Adam's Thesis isn't probably of *truth*.

[21] I say 'almost'—Gordon Belot is a persistent hold-out.

More strongly, it's hard to even come up with examples where Stalnaker's Thesis seems like it should be violated—at least if we restrict our attention to indicative conditionals.[22] It's less clear, however, that Stalnaker's Thesis gets things right in cases involving counterfactuals. To see this, first consider the following minimal pair (Bennett, 2003):

(10) a. If Shakespeare didn't write *Hamlet*, then someone else did.
b. If Shakespeare hadn't written *Hamlet*, then someone else would have.

Now, if I reflect on my own credences here—credences which I suspect are widely shared—then I find that I'm *extremely* confident that (10-a) is true. Moreover, it seems plausible that this is because my conditional credence that Hamlet was written by *someone*, given that it wasn't written by Shakespeare, is very high indeed (basically 1). Thus, my credence in (10-a) accords with Stalnaker's Thesis. However, the same isn't true of my credence in (10-b). On the contrary, in that case, my confidence is very low. And so it looks like, in the case of counterfactual conditionals, Stalnaker's thesis doesn't apply.

Stalnaker's Thesis, then, seems like the right precisification of the Ramsey test only when the conditionals in question are *indicatve* conditionals. At the same time, this doesn't imply that our judgments about the probabilities of counterfactuals are completely unconstrained. Quite the opposite, counterfactuals seem to obey an *analogue* of Stalnaker's Thesis, which was first described by Skyrms (1980a, 1980b), from whom we get the name:

> **Skyrms's Thesis**. Let ch_w be the *objective chance function* at world w (and at some contextually salient time). Then, your credence in the counterfactual conditional $A > C$ should be the following:
>
> $$p(A > C) = \sum_w p(w) \cdot ch_w(C \mid A).$$

This says that your credence in a counterfactual $A > C$ should be your expectation of the conditional objective chance of the consequent, given the antecedent. Once more, this principle seems to get things right in a broad range of cases. For example, consider again the sentence (10-b):

(10-b) If Shakespeare hadn't written Hamlet, then someone else would have.

Here, it seems like your credence in (10-b) should be low. And plausibly, that's because you know the conditional chance that someone writes *Hamlet*, conditional on Shakespeare not having written it, is very low as well. (Shakespeare, after all, was a singular genius.)

Thus, we have two precise ways of cashing out the Ramsey test. Specifically, our probability judgments about conditionals seem to be tied to conditional probabilities—conditional *credences* in the case of indicative conditionals, and conditional *chances* in the case of counterfactuals. If that's right, however, then we have another strong argument in favor of CEM (as I mentioned in the preceding section)—and *a fortiori*, an argument in favor of Stalnaker's semantics. To see this, consider the following derivation (I'll focus on Stalnaker's Thesis here; but much the same point could be made using Skyrms's Thesis). First, consider an arbitrary indicative conditional $A > C$. By Stalnaker's Thesis, we should have:

$$p(A > C) = p(C \mid A),$$

assuming that $p(A) > 0$. The same thing goes for $A > \neg C$. That is, by Stalnaker's Thesis, we should have:

$$p(A > \neg C) = p(\neg C \mid A).$$

[22] It's difficult, but not impossible. See, e.g., Kaufmann (2004) and Khoo (2016). I discuss the relevant examples in Chapters 3–4.

Now, by the probability calculus, it follows that $p(C \mid A) + p(\neg C \mid A) = 1$. But then, since $p(A > C) = p(C \mid A)$ and $p(A > \neg C) = p(\neg C \mid A)$, we have that:

$$p(A > C) + p(A > \neg C) = 1.$$

However, everyone agrees that $A > C$ and $A > \neg C$ are, at the very least, contraries. (CEM says something stronger, namely that they're contradictories. But I'm not assuming that here.) So it follows from another application of the probability calculus that:

$$p(A > C) + p(A > \neg C) = p((A > C) \vee (A > \neg C)).$$

Then, since $p(A > C) + p(A > \neg C) = 1$, we have that $p((A > C) \vee (A > \neg C)) = 1$. And this is an instance of CEM.

Thus, what we've established here is that, whenever your credences satisfy Stalnaker's Thesis, you're forced, on pain of irrationality, to assign the corresponding instance of CEM a credence of 1. This, in turn, provides a strong inductive argument in favor of CEM: our intuitive probability judgments about conditionals seem to support that principle.[23]

In contrast, theories of the conditional that *don't* validate CEM have a very difficult time accommodating probability judgments like the ones we saw above. To illustrate this, consider again the strict conditional theory, according to which a counterfactual $A > C$ is just a material conditional $\Box(A \supset C)$. In rough terms, the latter says that $\Box(A \supset C)$ is true at world w just in case all the worlds "accessible" from w are worlds at which $A \supset C$ is true. However, in the case of a sentence like (8) above, it seems plausible that not all Flip-worlds accessible from w will be Heads-worlds—after all, the coin is fair. Thus, the upshot is that, on the strict conditional theory, the sentence (8) comes out as *false*. And so, it seems like you should assign it a credence of 0.[24]

Once again, I think these sorts of probabilistic judgments about conditionals give us a very strong reason to favor Stalnaker's semantics. After all, as I said right at the outset, conditionals seem to play important roles in uncertain reasoning and deliberation. Thus, if other semantic views require a kind of *error theory* for our probability judgments about conditionals, that's a strong mark against those theories. According to them, conditionals can't play their natural roles in our epistemic and practical lives.

[23]Let me also note here that, even in cases where your credences *don't* seem to obey Stalnaker's Thesis, there's reason to think that corresponding instances of CEM should still hold. For example, consider the view put forward by Kaufmann (2004) and Khoo (2016), which says that your credence in an indicative conditional should sometimes abide by the following:

$$p(A > C) = \sum_i p(A > C \mid K_i) \cdot p(K_i),$$

where $\{K_i\}$ is a partition. (Note that this collapses to Stalnakers' thesis when the partition in question is the trivial partition.) If this rule is right, then notice that we have:

$$\begin{aligned} p((A > C) \vee (A > \neg C)) &= p(A > C) + p(A > \neg C) \\ &= \sum_i p(K_i) \cdot p(C \mid C \wedge K_i) + \sum_i p(K_i) \cdot p(\neg C \mid A \wedge K_i) \\ &= \sum_i p(K_i) \cdot p(C \mid A \wedge K_i) + p(\neg C \mid A \wedge K_i) \\ &= \sum_i p(K_i) \cdot 1 \\ &= 1. \end{aligned}$$

So, in this case, too, we're forced, on pain of irrationality, to assign the corresponding instance of CEM a credence of 1.

[24]Sarah Moss (MS) has recently objected to strict conditional theories in a similar way, and offered a new strict conditional theory—which she calls the *synthesis theory*—which is supposed to accord better with our intuitions about the probabilities of conditionals than extant strict conditional theories. See Moss (MS) for further discussion.

That said, we're now going to see that both Stalnaker's Thesis and Skyrms's Thesis face formidable challenges, having to do with probability. In particular, it turns out that if we accept either of those theses, then some intuitively plausible assumptions will have to go. If not, then these theses turn out to be *trivial* (in a technical sense, that I'll explain). And the upshot of *this* would be that the conditionals outlined in Stalnaker's theory can't play their requisite roles in our epistemic or practical lives either.

1.4 Triviality

The first blow to Stalnaker's Thesis involving probability was delivered by David Lewis (1976)—and it was bruising. In effect, Lewis showed that Stalnaker's Thesis is incompatible with some widely accepted background assumptions about conditionals and probability. In particular, Lewis's primary assumptions were that:

(i) Your credence function, p, is a classical probability function defined over a space of possible worlds, $\mathcal{W}$,

(ii) You update your credences by Bayesian *conditionalization*,[25] and

(iii) Your interptation of the sentence 'If A, then C' doesn't change when you learn new information.

At a first pass, all of these assumptions seem innocuous. The first two, for example, are part-and-parcel of the *Bayesian view* in epistemology (which I'm largely adopting here). Similarly, on the third point, Lewis says that:

> presumably our indicative conditional has a fixed interpretation, the same for speakers with different beliefs and for one speaker before and after a change in his beliefs. Else how are disagreements about a conditional possible, or changes of mind? (1976, p. 301)

That also seems right—at least initially. (Note, however, that later, I'll be rejecting this assumption of Lewis's. In fact, I'll be rejecting the first assumption, too.)

Now, unfortunately, if all of these assumptions are correct, then Stalnaker's Thesis (and, for that matter, Skyrms's Thesis, about which I'll say more in a moment) can't be. Given the assumptions (i)–(iii), we can show that Stalnaker's Thesis can hold for all probability functions in a class of putatively rational probability functions only if each such function deems the probabilities of A and C to be *independent*. In other words, for each such function in the class, we'd require that $p(A \wedge C) = p(A) \cdot p(C)$—or, equivalently (and more revealingly), that $p(C) = p(C \mid A)$. But this is clearly wrong. After all, just consider your intuitive probability judgment for this sentence:

(11) If I roll a prime number with this fair, six-sided die, then it will be an even number.

Here, it seems like your credence in (11) should be $1/3$. And by Stalnaker's Thesis, this implies that $p(\text{Even} \mid \text{Prime}) = 1/3$ as well (which seems right). But Lewis's results imply that $p(\text{Even}) = p(\text{Even} \mid \text{Prime})$. So, in turn, this implies that $p(\text{Even}) = 1/3$—which is clearly wrong.

After the publication of Lewis's landmark paper, a slew of results similar to his appeared in the philosophical literature, often making use of different, and weaker, assumptions than the ones Lewis himself

[25] Bayesian conditionalization says that, after learning a proposition A with certainty, your new credence in any proposition should be equal to your old credence in that proposition, conditional on A. That is, if p_A is your credence function after learning A:

$$p_A(-) = p(- \mid A),$$

provided that $p(A) > 0$.

made use of (see, e.g., Hájek, 1989, 2012; McGee, 1989; Hájek and Hall, 1994; Bradley, 2000; Fitelson, 2015; Khoo and Santorio, 2018; Goldstein and Santorio, 2021). Additionally, it wasn't long before similar results appeared affecting Skyrms's Thesis (Williams, 2012; Moss, 2013; Schultheis, 2023). So, neither of these theses seem like it can hold.

In Chapter 3 below, I've outlined Lewis triviality result for Stalnaker's Thesis—or rather, his *first* triviality result for that thesis—where I tie that result to another famous argument given by Lewis, against the so-called *Desire-as-Belief Thesis* (Lewis, 1988a, 1996). I've also described a different triviality result for Stalnaker's Thesis—this one due to Alan Hájek (1989)—in Chapter 4. Still, I think it's worth looking briefly at an example of such a triviality result at this early stage, to give you, the reader, a feeling for how they work. Thus, to avoid repeating myself, I'll describe in this section a clever triviality result for Stalnaker's Thesis due Goldstein and Santorio (2021). To see it, let's first note that Stalnaker's Thesis implies the following plausible principle:[26]

Positive Preservation. For all A, C such that $p(A) > 0$, if $p(C) = 1$, then $p(A > C) = 1$.

Positive Preservation seems completely obvious at a first pass. After all, consider an example adapted from Bradley (2000). Imagine that your credence that we'll go to the beach is positive, and that you're certain we'll go swimming even if we don't go to the beach. (Maybe, for instance, you think that we'll go to a pool if we don't end up going to the beach). Then, it would seem crazy for you to be less than certain of the conditional 'If we go to the beach, we'll go swimming'. For, as we just said, you're certain of the consequent!

We can show, however, that Positive Preservation leads to absurd results, in the presence of Lewis's assumptions (i)–(iii) above. To see this, consider the following derivation:[27]

$$\begin{aligned}
p(A > C) &\geq p((A > C) \wedge C) && \text{(Probability Theory)} \\
&= p(C) \cdot p(A > C \mid C) && \text{(Ratio Formula)} \\
&= p(C) \cdot p_C(A > C) && \text{(Conditionalization)} \\
&= p(C) \cdot 1 && \text{(Positive Preservation)} \\
&= p(C)
\end{aligned}$$

Thus, what this proof establishes is that, if you satisfy Positive Preservation—and *a fortiori*, if you satisfy Stalnaker's Thesis—then your credence in $A > C$ must always be *greater* than your credence in its consequent. But this clearly isn't right. For example, consider again the sentence (11):

(11) If I roll a prime number with this fair, six-sided die, then it will be an even number.

Once more, it seems like your credence in (11) should be $1/3$ (in line with Stalnaker's Thesis). But your credence in its consequent seems like it ought to be $1/2$. However, according to the triviality result we just proved, this distribution of credences is irrational. So, in consequence, this seems like a damning indictment of Stalnaker's Thesis.

1.5 Sequence Semantics

At this point in the dialectic, you'd be forgiven for thinking that both Stalnaker's Thesis and Skryms's Thesis are dead ends, and thus that one of my primary motivations for adopting Stalnakers's semantics in the

[26] To see this, note that if $p(A) > 0$ and $p(C) = 1$, then $p(C \mid A) = 1$. And by Stalnaker's Thesis, this implies that $p(A > C) = 1$.

[27] For a more explicit version of this derivation, see Goldstein and Santorio (2021). Note that, in the proof, Line 4 exploits the fact that, according to conditionalization, $p_C(C) = 1$.

first place—namely, that this semantics interacts nicely with our probability judgments—was misguided. Thankfully, however, there's a way of resisting the triviality results, both for Stalnaker's Thesis and for Skyrms's Thesis. And in this section I'm going to explore it. Resisting those results will require us to think about conditionals a little differently than we have been. But the reward for doing so will be a vindication of our probability judgments. Indeed, the construction I give below in some sense underpins all of the positive results that I give in the dissertation. So, I think its applications are very widespread.

Thus, to set things up, let's first return to my original gloss of Stalnaker's semantics. I said there that the sentence 'If A, then C' is true at a possible world w, according to this semantics, just in case the "minimally different" A-world to w is a C-world. After that, we attempted to capture the notion of "minimal difference" using a *selection function*. But this isn't the only way we could've done so.

In an important—but curiously neglected—paper, Bas van Fraassen (1976) showed how we can represent the notion of minimal difference, crucial to Stalnaker's semantics, in a slightly to different way. In particular—drawing on some insights from Lewis (1973b)—van Fraassen showed that if selection functions satisfy Stalnaker's constraints (i)–(iv), then this suffices to generate a *total ordering* of possible worlds, ordered according to how similar they are to a given world.[28] To illustrate this, suppose that $\mathcal{W} = \{w_1, w_2, w_3\}$. And imagine that w_1 is the possible world we're interested in. Then, Stalnaker's constraints on selection functions allow us to generate a *sequence* of possible worlds—say, for example, $\langle w_1, w_2, w_3\rangle$—which can be thought of as the hypothesis that w_1 is the minimally different world from itself, w_2 is the next most similar world, and so on.

Once we have sequences of worlds like this in place, we can state Stalnaker's semantics a little differently to how we did before. Very roughly: a conditional $A > C$ is true at a world w just in case the first A-world in the sequence of worlds beginning with w is a C-world. In essence, this is just another way of capturing the idea, inspired by Ramsey, that whether a conditional $A > C$ is true depends on what's true at a minimally different A-world. A possible world, after all, is the "ontological analogue of a stock of hypothetical beliefs".

You'll notice here that I said '*the* sequence of worlds beginning with w'. However, you'll also notice that, contrary to this definite description, there are usually many possible sequences of worlds that we can generate, given the choice of a first world, w. For example, even in the toy case, where $\mathcal{W} = \{w_1, w_2, w_3\}$, there are two sequences of worlds for each choice of the first world:

$$\langle w_1, w_2, w_3\rangle, \langle w_1, w_3, w_2\rangle,$$
$$\langle w_2, w_1, w_3\rangle, \langle w_2, w_3, w_1\rangle,$$
$$\langle w_3, w_1, w_2\rangle, \langle w_3, w_2, w_1\rangle.$$

So, which of these sequences gives the *true* notion of minimal difference?

This is where the contextualist aspect of my view kicks in. Recall that, in §1.2, I said that Stalnaker thinks there isn't a *unique* notion of minimal difference, suitable for all contexts. Instead, exactly what we *mean* by 'minimally different A-world' can change with context, depending on "the subject matter of the antecedent and consequent [of the relevant conditional], on features of the local conversational context such as the presumed interests of the participants of the conversation", etc. (Stalnaker, MS, p. 7). To illustrate this, go back to Jackson's example, which we considered in §1.2. The sentence there was:

[28] This is *almost* true. The main distortion is just that the model I build here—while inspired by van Fraassen—is closer to a model developed by Khoo and Santorio (2018), Goldstein and Santorio (2021), Khoo (2022) and Santorio (2022). The main difference is just, unlike in van Fraassen's model, the sequences of worlds I appeal to here are *finite* sequences, whereas in van Fraassen's model, they're infinite sequences. Nothing much turns on this—for the most part, I've adopted the "finitist" construction just because I find the mathematics easier to work with. Moreover, it's relatively straightforward to translate between the two models. That said, the model I develop here implies a slightly stronger background logic than the one Stalnaker himself makes use of (which is why I say that my semantics is *inspired by* Stalnaker's). But again, nothing much turns on this, for present purposes. See Mandelkern (forthcoming) for further discussion.

(6) If Fred had jumped off the roof of the Empire State Building, he would've died.

Now, when we considered this sentence initially, we said that it looked like the minimally different "Fred jumps"-world to actuality should be one where there's no net below Fred at the time of his jump, where gravity works as normal, and so on. But now suppose that you utter the following sentences, immediately after I've said (6): "You know, Fred doesn't have a death wish. He would've jumped off the roof only if there were a net in place to catch him. So,

(12) If Fred had jumped, he would've lived.

In this case, sentence (12) *also* looks true, in addition to sentence (6). But then again, it's clear that (6) and (12) can't be true together. The most obvious way to reconcile these judgments appeals to context-sensitivity. In particular, in the two different contexts, different notions of "minimal difference" are operating, and this explains why (6) can be true in one context, and (12) can be true in the other. More precisely, in the default context that obtains when (6) is uttered, the minimally different "Fred jumps"-world is (as I said) one where there's no net below Fred at the time of his jump. In contrast, the sentence you uttered in the run-up to (12) set up a new context, in which the minimally different "Fred jumps"-world *is* a world there's a net below Fred.

Thus, the appeal to context-sensitivity gives us a way of constraining the relations of minimal difference that are appropriate in a given context. At the same time, however, it doesn't solve all our problems. To see why, return to the coin flip example from §1.2. In particular, consider again the following sentences, which Lewis thought raised a problem for CEM:

(8) If I had flipped the fair coin, it would've landed heads.

(9) If I had flipped the fair coin, it would've landed tails.

Now, earlier, we heard that Lewis objected to CEM on the grounds that it's metaphysically implausible that one or the other of these sentences is determinately true at each possible world. And if we think that context always selects a *unique* sequence of possible worlds against which we assess conditional sentences, then Lewis's complaint here would be right. Recall also, however, that earlier I remarked on Stalnaker's reply to Lewis, according to which, in many contexts, (8) and (9) are both *indeterminate* in truth-value (even if their disjunction is true). In the present setting, the way we'll cash this out is by saying that there *isn't* a unique sequence of possible worlds against which we assess conditionals in such a context. Instead, we have to allow a multitude of sequences to count as "admissible"—some of which make (8) true, and others of which make it false.

This is exactly how Stalnaker thinks about things, given Lewis's objection—although he describes things in terms of selection functions, rather than sequences. In a 2021 paper, for example, he says that in most contexts:

> There will be many admissible selection functions… [I]n application, the context in which a conditional of any kind is interpreted may not fully determine the parameter relative to which the formal semantics specifies truth-conditions for the conditional. The idealized semantics makes a uniqueness assumption: for each proposition, there is a unique possible world (or possible situation) that is the possible situation in which the proposition is true, and that is minimally different from the actual situation (or more generally, the situation relative to which the conditional is being evaluated). But in practice, the relevant context may provide only constraints on the parameter that do not fully determine it. (pp. 102-03)

Translating this to sequences, we can respond to Lewis's complaint by saying that context usually makes a range of sequences admissible.

This, it turns out, is really the heart of van Fraassen's proposed response to Lewis's triviality results. In particular, he shows that, once we've allowed that there can be multiple admissible notions of "minimal difference" in a context, we can build a model in which Stalnaker's thesis is satisfied—and satisfied non-trivially. Building this model requires that we make some substantive, but reasonable, assumptions, especially when it comes to credences. But as I said, the reward for making these assumptions is a vindication of our probability judgments.

To see how things work, then, let's first give a slightly updated version of Stalnaker's semantics. Once more, recall that earlier, I said that a conditional $A > C$ is true at a world, w, on this semantics, just in case the minimally different A-world to w is a C-world. That definition is fine when there's just *one* relation of minimal difference that's admissible in the context. But as we just heard, many contexts seem like they're not like that. So, the definition we gave above needs to be changed to reflect this. In particular, if we were working with selection functions, then we'd now need to introduce a selection function as a parameter in our official semantics:

Stalnaker Semantics. $[\![\text{If } A, \text{ then } C]\!]^{w,f} = 1$ iff $f(A, w) \in C$.

(Once again, I should include a context variable, c, in the statement of this semantics. But I'll omit that variable here, to keep the notion uncluttered.) Thus, unlike our original statement of Stalnaker Semantics, this says that the point of evaluation for a conditional sentence isn't merely a world. Rather, it's a *pair* consisting of a world and a selection function.

Now, since selection function/world pairs correspond to sequences, we can re-state the semantic entry directly in terms of sequences, rather than world/selection function pairs. In particular, if s is a sequence of epistemtically possible worlds, then:

Stalnaker Semantics. $[\![\text{If } A, \text{ then } C]\!]^{s} = 1$ iff the first A-world in s is a C-world.

On the present semantic view, then, the points of evaluation for conditionals are *sequences*, rather than possible worlds. And this, as we just saw, is equivalent to the idea that the points of evaluation for conditionals are world/selection function pairs.

(Notice also that, since our semantics says that $A > A$ is true at a sequence, s, just in case the first A-world in s is an A-world, this is equivalent to saying that ordinary "factual" propositions correspond to sets of worlds. In particular, on the sequence-based view I've just outlined, a factual proposition A is true at s just in case it's true at the *first* world in s. The upshot is that, as we're currently thinking about them, worlds are entities which pin down the truth-values for all the "factual" propositions. They just don't pin down the truth-values for conditional propositions—those propositions are true or false at more fine-grained possibilities, namely sequences of worlds.)

Now, once we have this fine-grained view of conditional contents in play, it becomes clear that we need a way of extending your credence function, p, so that it's defined over sequences. After all, conditionals *look* like the type of thing that you can have probabilistic opinions about—otherwise why would Stalnaker's Thesis, or Skyrms's Thesis, be interesting? However, if your credence function is defined only over possible worlds, then it's unclear how we can capture these probabilistic opinions. The reason is that conditionals need no longer be true or false at possible worlds. But at the same time, your credence in a conditional is your expectation of its *truth*.

Thus, to make this extension, van Fraassen proposes that we proceed roughly as follows. First, starting with your original probability function, p, we "lift" this function to a new credence function, q, over sequences, using a recursive procedure (below, I write '$[w_1, ..., w_k]$' for the set of sequences beginning with $w_1, ..., w_k$ in that order):

(i) $q([w]) = p(w)$,

(ii) $q([w_1, ..., w_k]) = p(w_k \mid \mathcal{W} - \{w_1, ..., w_{k-1}\})$.

We can think of this recursive procedure in the following way. Roughly, it says that your credence in a sequence of worlds is just the probability, according to p, that you'd draw those worlds from an urn, in that order and without replacement. For example, your credence in the sequence $\langle w_1, w_2, w_3 \rangle$ in the toy example is just your credence that you'd draw w_1 from an urn first, followed by your credence that you'd draw w_2 next, having already drawn w_1, and so on. This seems like a very natural way of extending your credence function, p, to a function, q, defined over sequences. In essence, it says that the credences you assign to sequences are "parasitic" on the credences you assign to worlds.

There are a couple of important things to note about this "lifting" procedure. The first is that, because the credence function q preserves the credences that p assigns to possible worlds—that's more-or-less what clause (i) says above—it follows that q also preserves the credences that p assigns to "factual" propositions. Then, in light of this (since conditional probabilities are just ratios of unconditional probabilities), it follows that q preserves the *conditional* probabilities that the function p assigns as well. In a proper sense, then, q is an *extension* of p. It encodes all the information that p encodes, but more as well.

The more important thing to note, however, is that, when *all* the sequences of possible worlds are admissible in a context, van Fraassen shows that the extended credence function q satisfies Stalnaker's thesis. That is, he proves the following important result: [van Fraassen, 1976; Goldstein and Santorio, 2021; Khoo, 2022] Let q be probability function that extends p, according to the recursive procedure (i) and (ii). Let all sequences of possible worlds be admissible. Then,

$$p(A > C) = p(C \mid A).$$

That is, Stalnaker's thesis holds. This is a very striking result. Contrary to what Lewis's triviality results purported to demonstrate, van Fraassen's result shows that Stalnaker's thesis can hold non-trivially after all. To do so, we only have to allow that conditionals correspond to more fine-grained possibilities than just sets of worlds. And as we saw, this way of thinking fits very naturally with Stalnaker's original view. I've outlined how van Fraassen's "tenability result" works in more detail, in subsequent chapters. But the good news, for now, is that the statement of Theorem (12) above shows the triviality results are no longer so worrisome.

At the same time, van Fraassen's formal framework gives us the resources to explain why Stalnaker's thesis sometimes intuitively fails. Although I haven't discussed examples like this in this chapter, you can find specific cases to illustrate it in Chapters 3 and 4 below. The rough explanation for what's going on in those cases is that, in them, not *all* sequences of worlds count as admissible in the context. Instead, only a subset of sequences count as admissible. And this explains why Stalnaker's Thesis gets intuitively violated.

Let me also note here that, in recent years, versions of van Fraassen's "tenability result" have been proved for Skyrms's Thesis also. I myself have a version of such a result (McNamara, MS-c). And similar results have been given by Khoo (2022) and Schultheis (2023). Thus, the upshot is that, for anyone attracted to the probability judgments about conditionals which I've appealed to throughout this chapter, there are strong reasons to endorse the van Fraassen-type response to Lewis's triviality results. This response says that our intuitive probability judgments about conditionals can hold non-trivially. And that means that conditionals can play their natural roles in our epistemic and practical lives.

1.6 The Path Forward

Hopefully I've now done enough to convince you that the sequence-based semantics for conditionals, inspired by Stalnaker and van Fraassen, is a viable semantics, and that it has a number of impressive

benefits when it comes to the probabilities of conditionals. In the chapters that follow, I'll put this semantics to work to solve long-standing problems in decision theory and epistemology. (Or at least, *I* think those problems can be solved, using this semantics.) And this, I believe, functions as an indirect argument in its favor.

There is, however, still a lot of work left to be done on this semantics. For one thing, I've not yet got entirely clear on how we should think about the kind of indeterminacy that's involved in van Fraassen's construction.[29] Additionally, while van Fraassen's result shows that Stalnaker's Thesis can hold non-trivially—and the analogous tenability results show the same thing for Skyrms's Thesis—that's not yet quite as strong a result as we might have hoped. The reason is that both Stalnaker's Thesis and Skryms's Thesis are often interpreted as *normative* theses. That is, it's often said that, *if you're rational*, then your credences will satisfy these theses (at least in the appropriate contexts). The tenability results of van Fraassen and others, however, only establish that the theses *can* hold, not that they *should* hold. In particular, if your credences in sequences aren't parasitic on the credences you assign to worlds—as outlined by the recursive procedure (i)–(ii)–then Stalnaker's Thesis/Skryms's Thesis might be violated.

I've got some ideas for what we can say about these issues—and I've explored those ideas elsewhere, in other papers. For example, I've written about the issue of indeterminacy in a subsequent paper (McNamara, MS-c). And in co-authored work, Mikayla Kelley, Richard Roth, Snow Zhang, and I have attempted to tie some of van Fraassen's ideas to considerations of *accuracy* (MS).[30] Our hope is that, ultimately, an accuracy argument for Stalnaker's Thesis can be given, show that assigning credences in a way that violates this thesis leaves you worse off in terms of accuracy.

Those, however, are issues for another time. What I want to do now is turn to the open problems in decision theory and epistemology, which I think the semantics outlined here can help us to solve. The chapters to come are meant to be free-standing: you might think my solution is plausible in one case, for example, but implausible in another. (Additionally, since the chapters are free-standing papers, I often repeat myself in them, appealing to the same examples. I've also varied my notation across the chapters, depending on what seemed appropriate in the context.) Hopefully, however, you, the reader, will recognize the common theme that runs through the chapters. And hopefully, you'll see that, taken together, they function as an indirect argument in favor of the semantic view I've outlined here.

[29]Indeed, there's some reason to think that we don't *need* to appeal to indeterminacy here at all. For example, an alternative way we could go—one to which I'm occasionally sympathetic—is to a appeal to an "epistemicist" theory of vagueness, and say that, contrary to our intuitions, there really is a brute fact of the matter about which world is the unique, minimally different A-world in each context. That fact, however, isn't one that we can *know*. See Hawthorne (2005) for more on this idea.

[30]For more on the notion of accuracy, and how it fits into epistemology, see, e.g., J. M. Joyce (1999, 2009a) or Pettigrew (2016).

Chapter 2

Causal Decision Theory, Context, and Determinism

2.1 Introduction

Here is a bet—take it or leave it. You win \$1 if a proposition, P, is true, but you lose \$1 if P is false. Before you choose whether to accept or decline this bet, I'll tell you what P is. It's the proposition that the past state of the world, together with the laws of nature, determines that you accept.

Suppose you're certain of determinism. That is, suppose you're certain that the past state of the world, together with the laws of nature, determines whatever it is that you actually do (although in the present case, you're uncertain precisely *what* these things determine you'll do). Then, should you accept my bet? Or should you decline it? It seems perfectly clear that you should accept. After all, by your lights the proposition P is true only if you accept the bet. And it's false only if you decline. So, by accepting, it seems like you're sure to be a dollar better off than you'd otherwise be. Taking the bet is like accepting free money.

Cases similar to this one have come up quite often in the recent philosophical literature. And like the case just described, they're usually cases in which the best course of action is intuitively clear. Surprisingly, however, *causal decision theory* (CDT)—a theory that many regard as our best theory of rational decision-making—gets these cases wrong. It recommends courses of action that almost everyone can agree are irrational.

According to CDT, you should make choices by considering the expected *causal* consequences of your actions. Different versions of the theory attempt to make this idea precise in different ways. My preferred version—namely, the version of Stalnaker (1981b), refined by Gibbard and Harper (1978)—appeals to the close connection between causation, on the one hand, and *counterfactuals*, on the other. Roughly, it says that you should choose an option that you think *would* have a good outcome, *were* you to choose it.

However, the standard theory of counterfactuals—to which this version of CDT usually appeals—has a surprising upshot, if the laws of nature are deterministic. Specifically, it says that if anything, including the choice you make, were different in the present, either the laws would be violated or the distant past would be changed. It's this surprising upshot of the standard theory of counterfactuals that leads my preferred version of CDT to give the absurd recommendations in the cases that I mentioned. Other versions of CDT face similar difficulties, for closely related reasons.[1]

My aim here is to slightly refine the Stalnaker-Gibbard-Harper formulation of CDT, so that it avoids the problems posed by the "deterministic cases" I've been talking about. In my view, what these cases show

[1]See Skyrms (1980a, 1982, 1984), Lewis (1981), Sobel (1994), or J. M. Joyce (1999) for other versions of CDT. Then, see Ahmed (2013, 2014a, 2014b), Solomon (2021), Elga (2022), and Hedden (2023) for discussions of the problems raised by "deterministic cases" for these other theories.

isn't so much that there's a fault with CDT's guiding idea—that it's the expected causal consequences of your actions that matter for rational decision-making—but instead that Stalnaker-Gibbard-Harper CDT, at least as it's usually spelled out, doesn't pay sufficient attention to the *context-sensitivity* of counterfactuals. In response to this, I develop a "contextualist" version of Stalnaker-Gibbard-Harper CDT, which better accounts for this context-sensitivity. And I show that my theory avoids the problems faced by the classic formulation of CDT in determinstic worlds.[2]

In §2.2 below, I introduce the Stalnaker-Gibbard-Harper version of CDT, as well as the standard theory of counterfactuals. Then, in §2.3 I show that this theory gives the wrong recommendation in two well-known deterministic cases, both of which are due to Arif Ahmed (2013, 2014a, 2014b). In §§3.3–2.5 I introduce my theory: §3.3 starts with some background, as well as a general overview of the theory; and §2.5 gives some important further details. §4.4 then concludes the paper by returning to Ahmed's cases, and showing that my theory gets the right answer in them, as well as in related cases.

Before we get started, let me make two comments.

First, since nearly all of the cases I'm interested in here appeal to deterministic laws of nature, I'll assume determinism in what follows. More precisely, I'll assume that all the worlds under consideration obey deterministic laws. And I'll assume that this is something about which *you*—the agent facing the decision problems we discuss below—are certain. For present purposes, we can understand a system of laws to be deterministic just in case the following holds: any two worlds that obey those laws are either always exactly alike or never exactly alike, with respect to particular matters of fact (Lewis, 1979, p. 460). I'll leave it as a task for future work to see how well my theory generalizes to cases involving indeterministic laws. But for what it's worth, I think there's reason to be optimistic about its prospects.[3]

Secondly, some authors have recently argued that deterministic cases are not genuine decision problems. For, apparently, no agent who faces one can see herself as *free*.[4] This is something I disagree with. But for now I'll set my disagreement aside. Going forward, I'll assume that any agent facing a deterministic case can see herself as free, in some non-trivial sense. That my approach gets us the right answers in these cases, while also allowing us to make this assumption, is, I think, one of its main draws for those of us with both causalist and compatibilist commitments.

2.2 CDT and Counterfactuals

Whenever you face a choice, you'll have some *options* available to you, $A_1, ..., A_n$. Here, I'll take your options to be propositions, which—for now—I take to be sets of worlds. I'll also assume that your options form a *partition* of the space of worlds, in the sense that each world w is a member of exactly one A_i. Intuitively, we can think of your options as the finest-grained propositions you believe you can *make* true by deciding (cf. R. C. Jeffrey, 1983, p. 84).

You'll also have *outcomes* that can result from your choice, $O_1, ..., O_m$. I'll take these, too, to be propositions that form a partition. And I'll assume they're propositions whose truth would settle everything

[2]The approach I advocate for here is briefly suggested by Elga (2022, pp. 211-12) as an approach worth exploring. Also, while this paper was under review, I learned that Robert Stalnaker has recently sketched a response to a deterministic case that's broadly similar to mine (see §2.6.4, and his MS for details). There are a few important differences between Stalnaker's approach and mine, and I'll point these out as I go along. However, for the most part, I take this over-arching convergence to be good news: as the reader will notice, the view I spell out here is broadly Stalnakerian in spirit.

[3]A couple of other remarks about laws of nature. First, throughout, I use 'laws' and 'laws of nature' as a shorthands for 'fundamental physical laws of nature'. I also assume that laws of nature are inviolable. This assumption is not wholly uncontroversial (see, e.g., Lange (2000), Braddon-Mitchell (2001), and Kment (2006, 2014) for dissent). But I don't think rejecting it makes for a very promising response to the deterministic cases. So I won't explore it here.

[4]See especially J. M. Joyce (2016) and Solomon (MS). Note, however, that Joyce has stressed to me in conversation that he doesn't think being certain of determinism precludes the possibility that an agent can see herself as free *simpliciter*. Instead, he thinks this is merely a special feature of certain of the decision problems we'll encounter below.

that you care about.

Now, let cr be your credence function (subjective probability function). Let v be your subjective value function. And let $>$ be an operator, which takes a pair of propositions, P, Q, and returns the counterfactual $P > Q$. Then, CDT—at least in the Stalnaker-Gibbard-Harper formulation—says that you should choose an option, A, that maximizes *utility*, U, defined as follows:

$$U(A) = \sum_i cr(A > O_i) \cdot v(O_i). \tag{2.1}$$

As I said before, the idea here is that you should choose an option that you think *would* have a good outcome, *were* you to choose it.

Notice that I haven't yet mentioned causation. However, earlier, I said that, according to CDT, it's the expected *causal* consequences of your actions that matter for rational decision-making. So, we still need to say how the *counterfactual* rule above reflects this guiding idea. And to do that, we need to make some additional assumptions about the counterfactuals $A > O_i$.

For starters, let's assume they have the following standard semantics, due to Stalnaker (1968).[5] Let f be a *selection function*: a function that takes a proposition P and a world w as arguments, and returns a world $f(P, w)$, thought of, intuitively, as the "most similar" P-world to w. Then, Stalnaker's semantics says that a counterfactual $P > Q$ is true at w just in case Q is true at this most similar P-world, $f(P, w)$.[6]

Let's also make an assumption about the meaning of 'most similar P-world'. After all, not just any relation of similarity will do for present purposes. To see why, consider an example from Jackson (1977). Imagine that Fred is on the roof of a tall building, teetering on the edge. A moment later, he steps down. So I turn to you and say: "Thank goodness!

(1) If Fred had jumped, he would've died."

Puzzled by this, you respond to me: "That's not true; Fred's not suicidal. He would've jumped only if there had been a net below him. So,

(2) if Fred had jumped, he would've lived."

Here, it doesn't seem like either of us has said anything false. But then, it's also clear that the two counterfactuals we've uttered can't be true at the same time. The most plausible explanation of what's going on invokes *context-sensitivity*. When I uttered my counterfactual, we were in a context at which the most similar antecedent-world was one where there's no net below Fred at the time of his jump. When you uttered your counterfactual, we were in a context at which the most similar antecedent-world was one in which a specific causal precursor for Fred's jumping is salient—namely, there being a net below him. The function of your preamble—"That's not true; Fred's not suicidal…"—was to set up this latter context. Thus, my counterfactual is true in the first context, and your counterfactual is true in the second.[7]

Lewis (1979) calls counterfactuals like mine "standard counterfactuals", and counterfactuals like yours "backtracking counterfactuals". Very roughly, we can think of the former as counterfactuals for which the

[5] See also Stalnaker and Thomason (1970). Lewis (1973b) gives a very similar semantics for counterfactuals, although it differs from Stalnaker's in a few crucial ways. It's well known, however, that Lewis's semantics coincides with Stalnaker's, given the assumption of determinism. Thus, since I'm making that assumption in this paper, the differences between Stalnaker's theory and Lewis's aren't relevant here.

[6] This semantics assumes that there always *is* a P-world to be selected. A more general version of the semantics would relax this assumption, with a clause saying what happens when there's no P-world to be selected (see, e.g., Stalnaker (1968)). For present purposes, however, I'll set that case aside.

[7] I'm speaking loosely here. Really, it's the sentences that express counterfactuals that are context-sensitive, and not the counterfactuals themselves. But for present purposes, I'll mostly elide the distinction between propositions and sentences, since it simplifies things to do so.

most similar antecedent-world is one that's like the world of evaluation with respect to matters of fact in the past. And we can think of the latter as counterfactuals for which the past varies. (I'll revisit the former gloss later on.) Lewis also argues—convincingly, in my view—that it's only the first kind of counterfactual that can tell us about the *causal* effects of the antecedent on the consequent. And that, in a nutshell, is what we're after here. So, going forward, let's set backtracking counterfactuals aside, and assume that any counterfactual under discussion has a "standard" interpretation.[8]

To pin down the notion of a standard counterfactual more precisely, let's again follow Lewis—at least for now—in saying that, when P is about a nomically possible, dated event, the most similar P-world to w is one that's like w with respect to the following conditions:

(i) it matches w in all particular matters of fact at times before P, and

(ii) it obeys w's laws.

These criteria are plausible, not least because they deliver the right verdict in cases like Jackson's. To see this, just notice that, because there was no net below Fred when he was up on the roof, it follows by (i) that the most similar world at which he jumps is also a world where there's no net below him. Then, by (ii), it follows that Fred dies after jumping off the roof, since the most similar world at which he jumps is a world where gravity works the same as we're used to.

Notice also, however, that if w is a world with deterministic laws of nature, and P is a proposition that's false at w, then the most similar P-world to w can't be a world that satisfies (i) and (ii) perfectly.[9] After all, if the laws are deterministic, then the intrinsic state of the world at any time, together with the laws, determines its state at all times. Thus, if the most similar P-world to w matched w perfectly with respect to both (i) and (ii), it would have to be a world at which $\neg P$ is true. But by assumption, it's a world at which P is true. So at this world, a contradiction is true. And this makes P *counterfactually impossible*.

Since we're interested in spelling out CDT using counterfactuals, this isn't a consequence we can live with. So, we need to reject the claim that the most similar P-world to w is one that satisfies (i) and (ii) *perfectly*. Instead, we need to say something like: the most similar P-world to w is a world that provides the best *trade-off* between (i) and (ii).

The most influential account of this trade-off is, again, given by Lewis (1979). According to him, the best trade-off-world is one that matches w with respect to all matters of particular fact up until a time shortly before P, but which does not obey w's laws. Instead, it obeys a system of laws similar to those that obtain at w, but which permit a "local divergence miracle"—a small violation of w's laws, sufficient to bring P about.[10]

There are other ways we could go with respect to this trade-off, if we wished. For instance, Dorr (2016) gives a different account of similarity, according to which the best trade-off world is one that obeys w's laws perfectly throughout all time, and which is also like w with respect to "macro-history", but not with respect to "micro-history".[11] However, since causal decision theorists almost always work with Lewis's account

[8]Some philosophers argue that the distinction between standard and backtracking counterfactuals is merely one of degree, rather than kind (see, e.g., Holguín and Teitel (MS)). To make things simple here, however, I'm going to assume there's a clear-cut distinction between these two kinds of counterfactuals. For a well worked-out theory of this distinction, with which I'm broadly sympathetic, see Khoo (2017, 2022).

[9]The argument I give here closely follows Dorr (2016). Note that there's an unstated closure premise in the argument, as I state it. See Dorr's paper for a more careful presentation.

[10]See also Jackson (1977), Bennett (2003), Lange (2000), Kment (2006, 2014), and Khoo (2022).

[11]See Nute (1980), Bennett (1984), Albert (2000), Loewer (2007), Maudlin (2007), and Goodman (2014) for related accounts of similarity. Ahmed (2013, 2014b) denies that CDT can be underwritten by Dorr's account of similarity. But see Dorr (2016, §7) for a reply.

by default;[12] and since none of my conclusions would change if we adopted Dorr's account instead;[13] I'll take the former as my foil in what follows. From here on out, I'll call it the *miracles account*.

As an example of how CDT works when combined with the miracles account of similarity, consider the following decision problem (Nozick, 1969):

> *Newcomb.* In front of you are two boxes, A and B. Box A is opaque, and contains $\$1,000,000$ ($\$1m$) or nothing, but you don't know which. Box B is transparent, and contains a $\$1,000$ bill ($\$1k$). You have two options: either take just the opaque box (*One-box*); or take both boxes (*Two-box*). The catch is that, yesterday, a highly reliable predictor predicted which of these things you'd do. If she predicted that you'd take just the opaque box, then she put the million dollars inside that box. If she predicted that you'd take both boxes, then she left the opaque box empty. What is your choice?

Here's a table, representing your decision problem. (Note that here and throughout, I assume you value dollars linearly, so that $v(\$i) = i$, for any i.)

	Million	*No Million*
One-box	$\$1m$	$\$0$
Two-box	$\$1m + 1k$	$\$1k$

Table 2.1: *Newcomb*

Causal decision theorists all agree that you should take both boxes in *Newcomb*. After all, while there's a strong correlation between your choice and the predictor's prediction, that prediction is in the past and there's nothing you can do to change it. So, taking both boxes *causes* you to be better off, no matter what the predictor predicted.

To see that the version of CDT I sketched above delivers this verdict, notice that, no matter what you choose to do, the contents of the opaque box *would* be unchanged at the most similar world at which you chose differently, by the miracles account of similarity. Thus, taking both boxes gets you a thousand dollars more than taking one box *would*, no matter what the predictor put in the opaque box.

I won't go through the formal details of this argument, because the case is well known, and also because I'll be returning to it in §4.4 anyway. But the nice thing about mentioning the *Newcomb* problem now is that it illustrates a principle that's at the heart of CDT—the so-called *causal dominance principle*. According to this principle, if you're sure that one option will *cause* you to be better off than another, no matter what the world turns out to be like, then you shouldn't choose the latter option. This principle seems compelling. And it's ultimately what leads CDT to give (what I and many others think is) the right answer in *Newcomb*.

2.3 Deterministic Cases

CDT gets the right answer in *Newcomb*. But it gets the wrong answer in both of Ahmed's deterministic cases. In this section, I'll briefly review those cases, and spell out the answer that CDT gives in them.

One quick thing, before we get started. In both of the cases that follow, I assume there's a proposition, L, saying that some particular deterministic regularities are the (exceptionless) laws of nature. I also assume that you're almost certain this proposition is true (so, your credence in L is just a little short of 1). As we'll see, this assumption plays a special role in both of the cases to come.

[12] See, e.g., Gibbard and Harper (1978, p. 127, and pp. 160-61, n.2), Lewis (1981, p. 22, especially fn. 16), Sobel (1994, p. 42-43), and J. M. Joyce (1999, pp. 169-70).

[13] See, e.g., T. L. Williamson and Sandgren (forthcoming), Gallow (2022), Hedden (2023), and Kment (2023) for discussion of deterministic counterexamples that affect a version of CDT which makes use of Dorr's account of similarity.

2.3.1 *Betting on the Laws*

Here is the first case (Ahmed, 2013, 2014a):

> *Betting on the Laws.* You have a choice between two bets, and you must choose one of them. First, there's B_1, which pays \$1 if L is true, but pays nothing if L is false. Second, there's B_2, which pays nothing if L is true, but pays \$1 if L is false. You're certain that nothing you do can causally affect L's truth-value. You care only about winning the dollar.

	L	$\neg L$
B_1	\$1	\$0
B_2	\$0	\$1

Table 2.2: *Betting on the Laws*

Here, it seems intuitively clear that you should choose the first bet, B_1.[14] After all, you're almost certain of L's truth. And you're certain that nothing you do can causally affect its truth. Still, CDT says that it's permissible to choose the second bet, B_2. In other words, this theory says it's permissible to bet *against* your own credences.

To see why, recall the miracles account of similarity: if P is a proposition that's false at w, then the most similar P-world to w is one that matches w with respect to all matters of particular fact until shortly before P, but which does not obey w's laws. Now, with that in mind, suppose that L is actually true and you actually choose B_1. Then, happily, you win a dollar. But the miracles account says that, if you had chosen B_2 instead, you'd still have won a dollar, since the most similar B_2-world to actuality is one at which the proposition L is false.

Having reasoned your way to this conclusion, you should be certain of the following material conditional:[15]

$$(B_1 > \$1) \supset (B_2 > \$1).$$

The laws of probability then require that your credences satisfy this inequality:

$$cr(B_1 > \$1) \leq cr(B_2 > \$1).$$

Now we can plug these credences into CDT's equation (2.1):

$$\begin{aligned} U(B_1) &= cr(B_1 > \$1) \cdot 1 + cr(B_1 > \$0) \cdot 0 \\ &= cr(B_1 > \$1) \\ U(B_2) &= cr(B_2 > \$0) \cdot 0 + cr(B_2 > \$1) \cdot 1 \\ &= cr(B_2 > \$1). \end{aligned}$$

And since $cr(B_1 > \$1) \leq cr(B_2 > \$1)$, it follows that $U(B_1) \leq U(B_2)$. So, CDT says that choosing B_2 is rationally permissible. In fact, the theory says that choosing B_2 is rationally *required*, if you give any credence at all to the claim that L would be false no matter what you do.

Thus, what *Betting on the Laws* shows is that, sometimes, CDT tells you it's permissible to bet against the truth of a proposition in which you're almost certain, and whose truth-value you think is outside of your causal control. To my mind, however, no plausible decision theory ever says this. So *Betting on the Laws* is a counterexample to CDT, as we've spelled it out so far.

[14] Ahmed (2013, pp. 291-92) gives a formal argument for this claim, based on a principle he calls the *causal betting principle*. I think the intuition elicited by the case is sufficient for my purposes.

[15] Cf. Ahmed (2013, pp. 294–96).

2.3.2 *Betting on the Past*

Let's now consider Ahmed's second case (2014a, 2014b). It's a bit like the case we considered at the outset:

> *Betting on the Past.* In my pocket, I have a slip of paper on which is written a proposition H. I'm going to offer you two bets, and you must choose one of them. First, there's B_3, which pays \$1 if H is true, but costs \$10 if H is false. Second, there's B_4, which pays \$10 if H is true, and costs only \$1 if H is false. Before you choose between these bets, let me tell you what H is. It's a proposition about the intrinsic state of the world at some particular time in the distant past. Furthermore, you're certain that the truth of H, together with the the truth of L, determines that you accept B_3. And you're certain that the truth of $\neg H$, together with the truth of L, determines that you accept B_4.

	H	$\neg H$
B_3	\$1	$-\$10$
B_4	\$10	$-\$1$

Table 2.3: *Betting on the Past*

In this case, there's a compelling argument for the claim that you should choose B_3 (Ahmed, 2014a, p. 676; 2014b, p. 127). This is: you're almost certain that if you accept that bet, then you were determined by H and L to do so, and thus you're sure to win \$1. Conversely, you're almost certain that if you accept B_4 instead, then you were determined by $\neg H$ and L to do *that*, and thus you're sure to lose \$1. What better reason could you have for choosing B_3?

Still, CDT tells you to choose B_4, instead of B_3. The argument showing this is similar to the one we considered in the previous subsection.[16] To see how it works, first suppose that H is actually true. Then, by the miracles account of similarity, the most similar world at which you choose otherwise than you actually do is also a world at which H is true, since H is a proposition about the past. Thus, you should be certain of this material biconditional:

$$(B_3 > \$1) \equiv (B_4 > \$10).$$

By parallel reasoning, you should be certain of this material biconditional, too:

$$(B_3 > -\$10) \equiv (B_4 > -\$1).$$

The laws of probability then require that your credences satisfy these equalities:

$$\begin{aligned} cr(B_3 > \$1) &= cr(B_4 > \$10), \\ cr(B_3 > -\$10) &= cr(B_4 > -\$1). \end{aligned}$$

Now we can plug these credences into CDT's equation (2.1):

$$\begin{aligned} U(B_3) &= cr(B_3 > \$1) \cdot 1 + cr(B_3 > -\$10) \cdot -10 \\ &= cr(B_3 > \$1) - cr(B_3 > -\$10) \cdot 10 \\ U(B_4) &= cr(B_4 > \$10) \cdot 10 + cr(B_4 > -\$1) \cdot -1 \\ &= cr(B_3 > \$1) \cdot 10 - cr(B_3 > -\$10). \end{aligned}$$

[16] Cf. Ahmed (2014a, pp. 674-75).

As you'll see, $U(B_3) < U(B_4)$, no matter what your credences in the various counterfactuals. So, by CDT's causal dominance principle, it seems like you should choose B_4.

But this seems absurd. As Kment (2023, p. 7) remarks, for example, choosing B_4 seems hopelessly self-undermining: it's a bit like taking a bet on the claim that you don't accept that very bet. The upshot is that this case, too, is a counterexample to CDT, as it's currently been spelled out.

That said, not everyone agrees. Instead, some philosophers are willing to bite the bullet in *Betting on the Past*, because they think this case is relevantly similar to the *Newcomb* problem (and choosing *Two-box* in that case is something like a fixed point for causal decision theorists). As Elga (2022) says, for instance:

> In a standard Newcomb problem there is a causal dominance argument for taking two boxes: 'The \$1 million is either there or it is not, and you have no causal influence on whether it is. Either way (and no matter what else is true), taking two boxes gets you a better outcome than taking just one. So you should take two boxes.'... These conditions are satisfied in [*Betting on the Past*] just as much as they are in a standard Newcomb problem. So those who are sympathetic to the spirit of causal decision theory are under some pressure to endorse [taking B_4 in *Betting on the Past*]. (p. 207)

I disagree. To me, it seems like there are important differences between *Newcomb* and *Betting on the Past*. And in the next section, I'm going to begin spelling out a theory which, I think, helps us to see those differences.

2.4 Counterfactuals, Context, and Causation

Everyone—including Elga—agrees that at least one of Ahmed's cases poses a problem for CDT. However, philosophers sympathetic to that theory are divided about how to respond. For example, some say that we should modify CDT's decision rule (Sandgren and Williamson, 2020; T. L. Williamson and Sandgren, forthcoming; Solomon, MS). Others say that we should adopt a new semantics for counterfactuals (Gallow, 2022). And others still say that we should abandon CDT, and embrace a new decision theory in its place (Hedden, 2023; Kment, 2023). For my part, I don't think any of these responses are right—but I don't have space to discuss them here. So what I'll do instead is begin spelling out the alternative response that I favor. In my view, what Ahmed's cases show isn't so much that CDT is wrong, or that we need a new semantics for counterfactuals; it's that, as we've spelled it out so far, Stalnaker-Gibbard-Harper CDT doesn't pay sufficient attention to the *context-sensitivity* of counterfactuals.

Later on, I'll say why I think this is the right response to Ahmed's cases. But first, let me say a bit more about context-sensitivity for counterfactuals in general. To start, consider these sentences from Lewis (1973b), attributed to Quine:

(3) A. If Caesar had fought in Korea, he would've used nuclear weapons.
(3) If Caesar had fought in Korea, he would've used catapults.

Here, both sentences seem to have a standard interpretation. So, we should be able to use our current account of similarity—the miracles account—to pin down their respective truth-values. But unlike in other cases that we've seen, I, at least, have a hard time seeing how the miracles account is supposed to apply here. For one thing, it's not obvious just how much of history we're supposed to hold fixed when we're assessing (3) and (3).

Worse, it seems like there are contexts in which (3) would be true, and (3) would not (e.g., contexts in which present-day military technology is salient). And it seems like there are contexts in which (3) would be true, and (3) would not (e.g., contexts in which the military technology available to Caesar in his own day is salient). But again, it's difficult to see how the miracles account can deliver these verdicts on its

own. Instead, it seems like special features of the context—which have nothing to do with history before the antecedent-time—have to be cited, if we're going to get the right predictions about these sentences in the contexts in which it would sound natural to utter them.[17]

Thus, ironically, Lewis's case makes trouble for the claim that the miracles account applies straightforwardly to all standard counterfactuals. Here's another case, with an even more striking upshot. This one is from Dorr (2016, p. 265), and it has a similar flavor to *Betting on the Laws*:

> Suppose that L is a simple, true, deterministic law and that Frank, a philosopher of physics, has devoted his career to defending the truth of L. He is having a public debate with Nancy, who maintains (wrongly) that there are isolated exceptions to certain generalizations that follow from L, so that L is false. [The miracles account implies that] if the circumstances of the debate had been different in any way whatsoever—for example, if someone had put a glass of water on Frank's lectern, or rudely interrupted his talk—then Nancy would have been right and Frank wrong. Thus [(4)] and [(5)] are true:

(4) If we had given Frank a glass of water, his whole career would have been devoted to a mistake.

(5) If you had told Frank that his whole career was devoted to a mistake, you would have been right.

As Dorr rightly says, however, both (4) and (5) seem clearly false in this context. So this case, too, looks like it makes trouble for the miracles account.[18]

As a final example—and one with an additional upshot, as we'll see—consider a case from Slote (1978), credited to Sidney Morgenbesser. Imagine I've just tossed a fair coin, which is genuinely indeterministic, and I've offered you a bet while it's spinning in the air. (For a moment, set aside our earlier assumption about the laws of nature being deterministic.) If the coin lands heads, you win $1; but if it lands tails, you lose $2. Now, suppose you decline the bet, and a moment later the coin lands heads. I say to you: "That's a shame.

(6) If you had accepted, you would've won."

Most people think that this counterfactual is true. But if it is, we must be holding fixed a specific fact about history after the time of the conditional's antecedent. And it's not obvious how to square that verdict with the miracles account, since that account says what we hold fixed when we're assessing standard counterfactuals are facts about the past.[19]

Thus, each of the examples we just looked at seems to have a similar upshot. They each seem to show that there are some contexts in which the miracles account doesn't get the right verdicts about standard counterfactuals. Instead, there are contexts in which that account looks insufficient, on its own, to pin down the truth-values for sentences which we nevertheless judge true or false. And there are other contexts in which it looks like the account gives the wrong results entirely.

[17] Lewis might have agreed with this. In his 1979, for example, he says that (3) and (3) are each "true under a resolution of vagueness [viz., context-sensitivity] appropriate to some contexts" (p. 457). Confusingly, however, he then immediately goes on to discuss the distinction between standard and backtracking counterfactuals. And as I say in the main text, (3) and (3) *both* seem to have a standard interpretation. For related discussion of these examples, see Kment (2006, p. 263, fn. 4) and Ichikawa (2011, pp. 292-93). In any case, even if Lewis would agree with my verdict about (3) and (3), that's already a major step towards the kind of contextualism about counterfactuals that I prefer.

[18] Does this case constitute an argument for Dorr's alternative account of similarity, according to which we hold the laws fixed when we're assessing standard counterfactuals? Not in my view (although see Dorr (2016, §6)). For one thing, it doesn't follow from the fact that there are *some* counterfactuals for which we naturally hold the laws fixed that *all* standard counterfactuals require us to do this. See Holguín and Teitel (MS) for further discussion.

[19] This is a bit of a simplification. But see Edgington (2003) and Kment (2006, §3) for more in-depth discussions of why cases like this one are problematic for the miracles account.

The general lesson we should take away from these examples, I believe, is that counterfactuals are more sensitive to context than we earlier made them seem. While it's true that they can be sorted into the standard and backtracking categories, this doesn't exhaust the range of ways in which counterfactuals can be influenced by context. On the contrary, even if we focus on standard counterfactuals alone, it still looks like there's room for variation between contexts about the denotation of 'most similar antecedent-world'.

In philosophy of language, this fact has been widely acknowledged for some time.[20] But causal decision theorists don't seem to have paid it much attention. This is unfortunate, since the fact arguably has important consequences for what we think about Ahmed's cases. After all, in both of those cases, the miracles account played a key role in deriving the absurd recommendations. But then, if that account sometimes makes bad predictions about standard counterfactuals, it's reasonable to suspect that it's *this* that's leading CDT astray. In a case like *Betting on the Laws*, for example, you'd naïvely think that CDT would tell you to bet on the truth of the proposition L, rather than against it. And it seems like it's only because our current version of CDT relies on the miracles account of similarity that it says you should do otherwise.

This, indeed, is why I think CDT goes wrong in Ahmed's cases. So what I'm going to do now is sketch a more "contextualist" view of similarity for standard counterfactuals, and then a new version of Stalnaker-Gibbard-Harper CDT to go along with it. First, however, let me note one last thing about the final example we looked at. This is that there's a natural explanation for *why* we hold the outcome of the coin flip fixed in our assessment of (6)—namely, that this outcome is *causally independent* of whether or not you accept the bet.[21] This explanation looks especially plausible when we contrast the sentence (6) with the following sentence:

(7) If I had flipped a different (fair, indeterministic) coin, you would've won the bet.

Unlike (6), most people think that this counterfactual is *not* true. And the most natural explanation for why is that, while in the first case your choice to accept or decline the bet is causally independent of the coin flip's outcome, in the second case my choice of which coin to flip is not causally independent of the outcome. In other words, in (7), but not in (6), there's a causal chain running from the event described by the counterfactual's antecedent to the event described by its consequent. And this is why we think (6) is true, but (7) is not.

Lewis, whose account of similarity we've been working with up until now, wouldn't have accepted this explanation, because he believed that causation could be *analyzed* as a relation of counterfactual dependence between distinct events.[22] (Incidentally, that's why causal notions were nowhere mentioned when I was sketching the miracles account initially, despite the fact that I said standard counterfactuals often tell us about *causal* effects of the antecedent on the consequent.) However, I think examples like this one show strongly that a counterfactual analysis of causation can't succeed. So in what follows, I'm going to assume that causal notions *can* inform the truth-conditions for counterfactuals. Doing so makes available to us some resources that we didn't have before. And as we'll see, the idea that causal notions can influence counterfactuals plays an important role in how I spell out my theory.

[20] Thanks here to an anonymous referee, who points out that all of the following authors reject the miracles account in favor of an account of similarity that's more "contextualist" (note, however, that none of these accounts are exactly like the one I'll give below, nor do they ultimately get applied to CDT): Stalnaker (1968, 1981a, 1984, 2021), Ichikawa (2011), Ippolito (2016), Steele and Sandgren (2020). Additionally, even some philosophers who are broadly sympathetic to the miracles account—like Kment (2006) and Khoo (2022)—are critical of the version I gave in §2.2, and opt for a more contextualist version instead. For additional criticisms of the miracles account, different to the ones I've given here, see Elga (2001) and Holguín and Teitel (MS).

[21] This observation is also made by Bennett (2003, chapter 15), Edgington (2003), and Kment (2006, 2014), among others. See those works for further discussion, as well as for related examples.

[22] See Lewis (1973a, 1986, 2000).

2.4.1 Questions in Context

The examples we just looked at show that similarity relations for standard counterfactuals depend, not just on facts about the world's history before the antecedent-time, but also on more "local" matters, like features of a context that happen to be salient. In the Lewis/Quine case, for instance, the sentence (3) would plausibly be true in a context in which the existence of nuclear weapons is salient, but not in a context in which it's not salient. Thus, an adequate account of similarity for standard counterfactuals should give more weight to these features of context than the account we previously had. And what I'm going to do now is spell out one way in which I think this can be accomplished. The account of similarity I'll sketch below isn't so much a *theory* of similarity for standard counterfactuals, as it is a set of constraints which I think any plausible such theory should satisfy. But as we'll see, even this rough-and-ready account of similarity is sufficient to get the right answers in tricky cases like the ones we've seen.[23]

To start off, note that in fields like semantics and philosophy of language, it's common to think of salient features of a context as being represented by salient *questions*.[24] These questions *foreground* the issues that are "live" in the context, and they *background* the issues that aren't live (Yalcin, 2016, p. 30). In the Slote/Morgenbesser case, for instance, we can think of the salient issue as being represented by the question *How did the coin land?*. Then, contributions to the conversation are deemed relevant, or appropriate, just in case they address that salient question. (Note, however, that we don't have to assume this question is ever explicitly spoken in the context; it may be merely implicit.)

To make this idea precise, let's introduce some formalism. Following Stalnaker (1978), let's first say that any context can be modeled by a set of worlds, $\mathcal{W}$, which we call the *context set*. For simplicity, I'll assume that $\mathcal{W}$ is always a finite set of worlds. And intuitively, we can think of it as the set consisting of all worlds that count as "live options" in the context, for the purposes at hand. For instance, in an ordinary conversational context, $\mathcal{W}$ might consist of all the worlds that you and other conversational participants believe could be actual. And in a deliberational context, $\mathcal{W}$ might consist just of your epistemically possible worlds.

Now, a salient question can be thought of as a *partition* of the context set.[25] Each cell of this partition groups together worlds that are alike with respect to a complete answer to the question. And any union of these cells corresponds to a partial answer. In the Slote/Morgenbesser case, for example, the partition is just the set which groups together *Heads*-worlds and *Tails*-worlds, respectively. Similarly, in Dorr's "Frank vs. Nancy" case, where the question *Is L a law of nature?* is salient, the associated partition consists of worlds that obey the L-law, and worlds that don't, respectively.

Questions like this give us a way of constraining the similarity relations that are appropriate in a context—or *admissible*, as I'll often say. Specifically, when there's some feature of the context that's especially salient, we can think of an admissible similarity relation as being one that "holds fixed" the answers to a corresponding question. To see what this means, let $\mathcal{W}$ again be a context set, and let $Z = \{Z_1, ..., Z_n\}$ be a partition of $\mathcal{W}$, corresponding to such a question. Then, in my view, a similarity relation is admissible in this context only if its associated selection function satisfies the following constraint: for each world $w \in \mathcal{W}$, if $w \in Z_i$, then $f(P, w) \in Z_i$, where P is the antecedent of the counterfactual of interest. In

[23] The contextualist view of similarity I sketch in this section has a lot in common with views espoused by, e.g., Kaufmann (2004), Ippolito (2016), Khoo (2016), Boylan and Schultheis (2021), and Dorr and Hawthorne (MS). It also has something in common with so-called *causal modeling* approaches to counterfactuals, like those of Hiddleston (2005), Santorio (2019), Gallow (2022), or Khoo (2022). See also J. M. Joyce (2009b).

[24] The *locus classicus* for this view is Roberts (2012). However, there's an important difference between the way I'm understanding the notion of a salient question—or a *question under discussion*, as Roberts calls it—and the way Roberts herself does. In particular, I'm not going to assume that these questions are always *unanswered* in a context. See Boylan and Schultheis (2021) for a similar understanding of salient questions.

[25] See Groenendijk and Stokhof (1984), and also Hamblin (1973). A different, but equally plausible, way to think about this partition is in terms of *subject matters*. See Lewis (1988b, 1988c). I'll stick with the notion of questions in the main text to streamline the discussion.

words: a similarity relation is admissible in a context only if it says that the most similar P-world to w is one that lies in the same cell of the salient partition as w itself. (See Figure 2.1 for an illustration.)

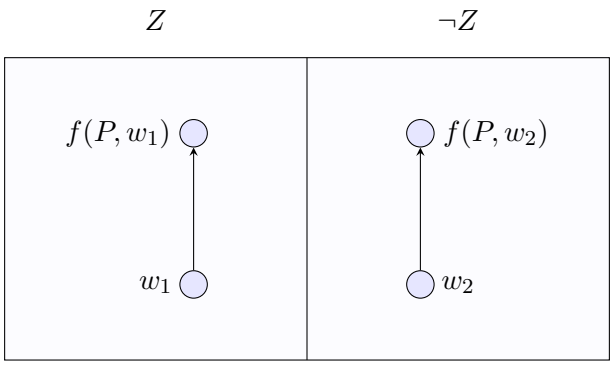

Figure 2.1: An Admissible Similarity Relation.

We can make this idea clearer by using a concrete example. So, consider again the Slote/Morgenbesser case, and particularly the sentence (6):

(6) If you had accepted, you would've won.

As I said before, the salient question here is *How did the coin land?*. And the associated partition consists of *Heads*-worlds and *Tails*-worlds, respectively. Now, at the actual world, we know that we're in the *Heads*-cell of this partition, since the coin actually landed on heads. And we know, further, that the actual world is one at which you didn't accept the bet. Thus, according to the account of similarity I'm sketching, the most similar world at which you *do* accept the bet is also a world in the *Heads*-cell of this partition, on any admissible similarity relation. The upshot is that the sentence (6) comes out true at the actual world. And this is the result for which we were hoping.

Already, then, this broad-brush account of similarity for standard counterfactuals gets a case right, which the miracles account got wrong. There are a few interesting things to note about it. First, it's consistent with the view as I've spelled it out so far that there can be more than one admissible similarity relation in a context. After all, if a similarity relation is admissible just in case it holds fixed the answers to a salient question, then in general there will be many such relations that can do this job. Thus, the account of similarity I'm sketching allows for some *indeterminacy* in the interpretation of counterfactuals. Specifically, when there's more than one similarity relation in play in a context, it will be indeterminate which world is picked out by the phrase 'most similar P-world'. When that's so, a sentence like 'If P, would Q' can be indeterminate in the context, since there will be some admissible similarity relations which make it true, and others which make it false. This fact will be important later. But for now, just note that it's in keeping with the idea that questions represent the distinctions we're interested in making in a context. In other words, since questions group together worlds according to aspects of similarity that are contextually salient, pinning down a similarity relation more precisely than this might give us more information than is relevant.

Additionally, note that nothing I've said so far rules out there being more than one salient question in a context. (We can, however, always assume there's *at least* one salient question in every context, since the trivial question, $\mathcal{W}$ itself, always counts as salient.) In particular, if one question *contains* another—in the sense that every cell of the partition corresponding to the second question is a union of cells of the partition corresponding to the first—then, if the more fine-grained question is salient in a context, the more coarse-grained question will be, too. I'll call the first question here a *refinement* of the second, and

the second a *coarsening* of the first.

In cases like this, we can usually think of similarity relations as being constrained by the *most* fine-grained question in a context. This goes also when we have two (or more) such questions, and neither is more fine-grained than the other. For example, suppose that in the Slote/Morgenbesser case, we were interested in the question *How did the coin land?*, but also in the question *Is the coin a nickel or a dime?*. Then, in that case, we could think of the similarity relations for counterfactuals as being constrained, not by either of these questions alone, but by their *conjunction*—or, more precisely, their *coarsest common refinement*. This is the partition each of whose cells corresponds to an intersection of cells from the first question and cells from the second question. For instance, the coarsest common refinement of *How did the coin land?* and *Is it a nickel or a dime?* is the partition: {*Heads and Nickel, Heads and Dime, Tails and Nickel, Tails and Dime*}.

This is useful to know about, but it won't play much of a role in what's to come. There is, however, another notion which will turn out to be important. In some contexts where there's more than one salient question in play, it's appropriate to think of similarity relations as being constrained, not by their coarsest common refinement, but by their *finest common coarsening*. This is the most fine-grained question that's coarser than both of the questions we started with. For instance, the finest common coarsening of *How did the coin land?* and *Is it a nickel or a dime?* is just the trivial partition, $\mathcal{W}$, since the only "question" that's coarser than both {*Heads, Tails*} and {*Nickel, Dime*} is the context set itself. The notion of a finest common coarsening of questions is a bit like the *disjunction* of those questions.[26] But it's a tricky notion to get your head around. So I'll defer further discussion of it until it's needed.

In the meantime, let me note one other constraint that needs to be imposed on questions-partitions, if my account of similarity is going to work as intended. It's important that this account not make counterfactuals vacuously true in a context. But for all I've said so far, nothing rules out this being the case: it may be that $Z = \{Z_1, ..., Z_n\}$ is the salient partition, but $P \cap Z_i = \varnothing$, for some Z_i and counterfactual antecedent P. In cases like this, a counterfactual beginning with P will be vacuously true at any world $w \in Z_i$, since there simply won't be any P-worlds in that cell in the first place. In general, however, when we assess counterfactual sentences in a context, we try to do so in ways that don't make them vacuously true. Thus, to capture this idea, I'm going to assume that, if P is a counterfactual's antecedent, then any suitable partition that could constrain relations of similarity for this counterfactual satisfies $P \cap Z_i \neq \varnothing$, for each Z_i. Khoo (2016) calls this the *well-definedness* constraint on question-partitions, and here I'll follow suit.

Now, if you go back and check the examples we've looked at so far, you'll see that it's often easy to spot a candidate question, and that, when similarity relations are constrained by this question, we get the right verdicts about the counterfactuals. In Dorr's case, for instance, the salient question corresponds to the partition $\{L, \neg L\}$. Then, since we're told that the actual world lies in the L-cell of this partition, the sentences (4) and (5) both come out false in this context. This, as I noted before, is the result that we were after.

But there's still one important way in which my account needs to be refined. To see what it is, consider again the sentence (7):

(7) If I had flipped a different (fair, indeterministic) coin, you would've won the bet.

Suppose we analyzed this sentence in the same way we analyzed (6). That is, suppose we took the salient question to be *How did the coin land?* And suppose we took the corresponding partition to consist just

[26] Arguably, it's not exactly like the disjunction, however. One reason is that the disjunction (union) of partitions need not be a partition. And if we think of questions as being modeled by partitions, then the disjunction of two questions might not itself be a well-formed question. This is plausibly related to the fact that, in natural language, disjunctions of questions often strike us as infelicitous. (For example: 'Did the coin land heads or tails, or is it a nickel or a dime?'.) As we'll see later on, I think some of this applies to Ahmed's *Betting on the Past* case.

of *Heads*-worlds and *Tails*-worlds. Then, since the actual world lies in the *Heads*-cell of this partition, it seems like any admissible similarity relation will say that the most similar world to actuality at which I flipped a different coin is a world at which you win the bet. The sentence (7) thus comes out true. But earlier, I said this sentence is *not* true.

The reason our analysis goes wrong here—as I tried to stress before—is that, in the case of (7), the coin flip's outcome is not causally independent of the counterfactual's antecedent. This immediately suggests one final constraint we need to impose on question-partitions. If a partition is going to constrain the admissible similarity relations for standard counterfactuals in a context, then it has to consist only of propositions that are *causally independent* of the antecedents. Without that constraint, our account will make bad predictions in cases like the one we've just seen.

With this constraint, however, the account gives plausible verdicts. The constraint also gives us a way of characterizing question-partitions for counterfactuals more generally—at least when the antecedents of those counterfactuals are about nomically possible, dated events. Often, we can think of the propositions in these partitions as being ones that describe salient "causal background factors", with which a counterfactual's antecedent would combine to *bring about* the consequent. For example, in the Slote/Morgenbesser case, every contextually-relevant world at which the coin lands heads is one at which taking my bet *causes* you to win \$1; and every contextually-relevant world at which the coin lands tails is one at which taking the bet causes you to lose \$2. Thus, how the coin lands is the only contextually-relevant background factor in this case. And that's why it makes sense to think of the relevant question-partition as holding only this background factor fixed. After all, as we heard before, standard counterfactuals often tell us about the *causal* effects of the antecedent on the consequent. So what we hold fixed when we're assessing these counterfactuals is often a contextually salient causal background, against which the counterfactual's antecedent takes place.

There's more to be said about this in general. But thankfully, in all the cases we'll consider here, it'll be straightforward to see what this causal background consists in.

2.4.2 Contextualist CDT: A First Pass

The account of similarity for standard counterfactuals I just sketched is more flexible than the miracles account with which we started. And this should give us some hope that a version of Stalnaker-Gibbard-Harper CDT, equipped with this account of similarity, can avoid the problems posed by Ahmed's cases. That said, there are a few things we still need to figure out. One of them is how we're supposed to think about the notion of a "salient question" in a context where you're making a decision, rather than having a conversation. This isn't yet obvious. But thankfully, it turns that there's a very natural way to think about these questions in decision-making contexts. Daniel Hoek (2019, 2022) has recently investigated this idea at length, and what I'll say here is broadly in line with his suggestions.[27] To see how the idea works, let's go back to the *Newcomb* problem:

[27]I should note, however, that there are a few important differences between my theory and the one that Hoek develops. For instance, Hoek (2019) seems to think that question-partitions are induced by similarity relations, rather than constraining those relations. Also, Hoek (2019, 2022) doesn't discuss the role of context in making certain questions salient. And most importantly, my theory allows that there can be multiple, competing questions "raised" in a decision-making situation, which is something that Hoek doesn't discuss. As we'll see, this fact also necessitates the distinctive formal apparatus that I introduce in §2.5. (This is also one of the ways my theory differs from the view of Stalnaker (MS).

	Million	*No Million*
One-box	$\$1m$	$\$0$
Two-box	$\$1m + 1k$	$\$1k$

[
Newcomb]Table 2.1: *Newcomb*

Here, the rows of the table correspond to your options, and the columns correspond to propositions about "states of the world", which (by your lights) may or may not obtain. The conjunction of any option with any state-proposition entails a unique outcome—that's why we can represent the decision problem using a table like this one. Notice, however, that if this representation is going to work, then it's important that the state-propositions form a partition. Thus, since we're thinking of questions as corresponding to partitions, there's a very natural candidate for the salient question in *Newcomb*. This is the question *How much money is in the opaque box?*, corresponding to the partition {*Million*, *No Million*}.[28]

Similar things can be said about the other decision problems we've looked at. That is, in each of the cases we've seen, there's a partition of states given in the corresponding decision table, with the features that (i) each of your options is consistent with each cell of this partition, and (ii) the conjunction of any state-proposition with any one of your options entails a unique outcome. Thus, it's natural to think of these partitions, too, as corresponding to salient questions. And in fact, whenever we have a partition of states in a decision problem with the features (i) and (ii), it seems—prima facie—like we can take that partition to correspond to a salient question. Then, we can use these partitions to fix the admissible similarity relations, in a way analogous to the way we saw before. That is, in parallel to what I said in the last subsection, a similarity relation is admissible in a context only if it says that the most similar world to w at which you choose some particular option lies in the same cell of the state-partition as w itself.

This works whenever the state-partition consists of propositions that are causally independent of your options. But then again, not every decision problem has this feature. To illustrate, consider a well-known case from J. M. Joyce (1999). Imagine that you've parked your car in a seedy neighborhood, when a man approaches you and offers to "protect" your car for the low fee of $\$10$. You know that people who don't pay the fee invariably come back to find their windshields smashed. And you know that any repairs to your windshield would cost you $\$100$. Now, the natural way to represent your decision problem is as follows:

	Smashed windshield	*Unsmashed windshield*
Pay	$-\$110$	$-\$10$
Don't pay	$-\$100$	$\$0$

Table 2.4: *The Shakedown*

But the salient question constraining the similarity relations in this context can't be the one corresponding to the partition {*Smashed windshield, Unsmashed windshield*}. If it were, then every admissible similarity relation would say that you do better by not paying than by paying. But that seems absurd: by not paying, you *cause* the man to smash your windshield. And by paying, you cause him to leave your windshield alone.

As before, then, we need to assume that any partition that constrains similarity relations for standard counterfactuals in a decision-making context is one which specifies only propositions whose truth-values are causally independent of your options. Sometimes, this will mean that the partitions which constrain these relations don't match up with the corresponding decision tables. There is, however, a general way in

[28]Actually, there are really two salient questions in *Newcomb* (and similarly for other decision problems). The second question corresponds to the partition of your options. However, because your "answer" to this question depends on what you believe about the other question, we can generally take the partition of states to be the *most* salient question in a decision problem.

which we can think about these partitions, analogous to what I said before. This is: usually, we can think of them as specifying salient "causal background factors" with which an option combines to cause a specific outcome.[29] In Joyce's case, for example, this partition might consist of the propositions *The man is a villain* and *The man is not a villain*, since, here, the man's temperament is the only salient causal background factor determining what the outcome of your options would be. Similarly, in the *Newcomb* problem, the only relevant causal background factor is whether or not the million dollars is in the opaque box. At every contextually relevant world, once you know whether or not the money's in that box, you know everything you need to know about what your choice of an option would cause.

Now, given this way of thinking about question-partitions in decision-making contexts, there's a generic, English language gloss we can give of these questions. This is: *How do the things I care about—viz., outcomes—depend causally on what I do?*. When it's put in these terms, the question might remind you of something. Both Skyrms (1980a) and Lewis (1981) give versions of CDT which appeal to propositions called *causal dependency hypotheses*. In Skyrms's words, these are "maximally specific specifications of the factors outside our [causal] influence at the time of decision, which are causally relevant to the outcomes of our actions" (p. 133).[30] This sounds pretty similar to the thing I'm now proposing.

Broadly speaking, this is right. But there are a few crucial differences between "dependency hypotheses", as I'm thinking about them, and the way that Lewis and Skyrms do. First and most obviously, Lewis thinks these propositions hold in virtue of patterns of counterfactual dependence, since his view is that causation just *is* a relation of counterfactual dependence between distinct events. Earlier, however, I said that my view is that (standard) counterfactuals often hold in virtue of causal relations. So there's a sense in which Lewis and I are approaching things from opposite directions. Whereas he thinks that counterfactuals come before dependency hypotheses in the order of explanation, I think that the reverse is true.

More importantly, unlike both Lewis and Skyrms, I'm not requiring dependency hypotheses to be *maximally specific* propositions. On the contrary, the view that they *are* maximally specific propositions looks untenable, if the laws of nature are deterministic. As Hedden (2023) points out, for example:

> something that doesn't causally depend on which of your present actions you perform can nonetheless entail which one you do. This means that if dependency hypotheses can specify anything that doesn't causally depend on your present action [like history and the laws of nature], then we'll have some dependency hypotheses which are inconsistent with some of your available actions, resulting in actions with undefined [utility]. (p. 744)

The upshot is that, if the laws of nature are deterministic, the theories of Lewis and Skyrms will simply fall silent in certain decision problems. This seems like an even worse problem than the ones we encountered in §2.3.

But like I said, I'm not requiring "dependency hypotheses" to be maximally specific propositions. All I'm requiring is that they specify salient causal background factors which, by your lights, are sufficient to cause outcomes, in conjunction with your choice. Which factors those are is a context-sensitive matter. And it's this sensitivity to context that allows me to avoid the problems faced by Lewis and Skyrms.

2.4.3 Loose Ends

Having now sketched most of the background for my theory, you can probably already tell how it's going to work in particular cases. Unfortunately, however, we're not yet in the clear. To see why, consider again

[29] If the laws are indeterministic, then we might need to say something more general here. After all, in that sort of case, even if we hold fixed the complete past and laws of nature—both of which are causally independent of your options—your choice might nevertheless only be sufficient to causally determine the *chance* of some outcome, rather than the outcome itself. Since I'm assuming determinism here, however, I'll ignore this possibility.

[30] Lewis (1981, p. 11) gives a gloss of 'dependency hypotheses' that's even more similar to the one I just gave.

the *Betting on the Past* case. Recall that the decision table I used to represent that decision problem was:

	H	$\neg H$
B_3	\$1	$-$\$10
B_4	\$10	$-$\$1

[
Betting on the Past]Table 2.3: *Betting on the Past*

And here, it's clear that the partition $\{H, \neg H\}$ corresponds to a salient question, in the sense I defined above. But notice: given what you know about the salient proposition L in *Betting on the Past*—namely, that it determines you choose B_3 in conjunction with H, and determines you choose B_4 in conjunction with $\neg H$—the following table also seems like a good representation of your decision problem.

	L	$H \wedge \neg L$	$\neg H \wedge \neg L$
B_3	\$1	\$1	$-$\$10
B_4	$-$\$1	\$10	$-$\$1

Table 2.5: *Betting on the Past*, Version 2

Indeed, even Ahmed acknowledges this. In his 2014a, for example, he says that "[Table 2.5] is just as accurate as [Table 2.3] when it comes to representing [your] situation. It represents the same payoffs to the same actions in the same circumstances at all the possible worlds where this could matter to a causalist" (p. 677).[31] If this *is* an adequate representation of your decision situation, however, then it looks like the partition $\{L, H \wedge \neg L, \neg H \wedge \neg L\}$ also counts as a salient question. And in *this* case, it's clear that the causal dominance argument for B_4 doesn't go through. Indeed, if utility is calculated in line with Table 2.5, then CDT will recommend B_3.

This is peculiar. What we seem to have is a case in which there's more than one salient question raised by your decision situation—something I earlier said was possible. But problematically, depending on which of the questions we focus on, CDT gives different recommendations about what you should do.[32]

[31]Actually, the version of the table that Ahmed considers in his 2014a is the following, since in that paper he assumes, not just that you're highly confident of L, but that you're certain of it.

	L	$\neg L$
B_3	\$1	$-$\$10
B_4	$-$\$1	\$10

My Table 2.5 is a bit more complicated, because, in the version of *Betting on the Past* I've given here, you give a tiny amount of credence to the possibility that L is false. Thus, for you, there some worlds which get positive where you choose B_4 and H is true. And there are some worlds that get positive credence where you choose B_3 and $\neg H$ is true. This is why I've fine-grained the $\neg L$-column in Ahmed's table.

[32]Now's a good time to note that we can't simply focus on the coarsest common refinement of the two partitions in this case either. After all, this coarsest common refinement corresponds to the following partition:

	$H \wedge L$	$H \wedge \neg L$	$\neg H \wedge L$	$\neg H \wedge \neg L$
B_3	\$1	\$1	$\varnothing$	$-$\$10
B_4	$\varnothing$	\$10	$-$\$1	$-$\$1

This fails my well-definedness condition on question-partitions. See J. M. Joyce (2016), Solomon (2021), Elga (2022), Hedden (2023), and Fusco (forthcoming) for further discussion of this table. You can probably see now why the notion of a finest common coarsening of questions is going to be important.

Some philosophers will respond to this by saying that *Betting on the Past* isn't a well-posed decision problem. Others will say that there isn't a univocal answer about which option you should choose, since CDT makes different recommendations depending on how the problem is represented. I won't pursue either of those responses (although I have some sympathy with the latter). My chief reservation about them is just that, in *Betting on the Past*, I have very clear intuitions about which option it's rational to choose: as I said before, choosing B_4 seems hopelessly self-undermining.

Besides, there's a more general problem looming here. To see what it is, recall from earlier that I said it's consistent with the contextualist view about similarity that I like that context can sometimes underdetermine which world counts as "most similar". This is a general feature of contextualist views about counterfactuals. But what *Betting on the Past* shows, I think, is that, when there's more than one question that's salient in a context, this kind of indeterminacy needn't be inert. Instead, it will occasionally lead CDT to give conflicting recommendations, depending on how 'most similar P-world' is precisified.

Thus, before I can state my own version of CDT completely, we need to find a way of handling this kind of indeterminacy. That's the task of the next section. And it's that task to which we now turn.

2.5 Accommodating Indeterminacy

Let's go back to Stalnaker's semantics. Recall that, when I introduced that semantics initially, I appealed to the notion of a selection function: a function from propositions and worlds to possible worlds. Now, it turns out that if we assume selection functions satisfy some natural constraints, then Stalnaker's semantics can be specified in a slightly different way. To see this, let me first state the constraints I have in mind. They are:

(i) **Success**. $f(P, w) \in P$.

(ii) **Strong Centering**. If $w \in P$, then $f(P, w) = w$.

(iii) **Reciprocity**. If $f(P, w) \in Q$ and $f(Q, w) \in P$, then $f(P, w) = f(Q, w)$.

(iv) **Accessibility**. If $w \in \mathcal{W}$, then $f(P, w) \in \mathcal{W}$.[33]

(Remember: $\mathcal{W}$ here is the context set.) Each of these constraints is very plausible. For example, Success just says that the most similar P-world to w should *be* a P-world—and that seems obviously right. Strong Centering says that, if $w \in P$, then w should count as the most similar P-world to itself—and that, too, seems right. Reciprocity is needed to validate a host of compelling inference patterns involving counterfactuals. And Accessibility says that selection functions shouldn't "reach outside" the set of worlds that are relevant in the context. A little reflection shows, additionally, that the first three constraints in particular are needed if selection functions are going to track anything like a *similarity* relation between possible worlds. And that, of course, is what we're after here.

Now, it turns out that the constraints (i)–(iv) above suffice to ensure that selection functions totally order the worlds in $\mathcal{W}$. That is, given the choice of a "base world", w, selection functions "rank" the worlds in $\mathcal{W}$ according to how similar they are to w, with w always counting as the most similar world to itself.[34]

[33]The fact that I'm introducing the Accessibility constraint here might surprise you, since this constraint is usually taken to characterize *indicative* conditionals, rather than counterfactuals (see, e.g., Stalnaker (1975)). I'll say more about why I'm introducing accessibility below. In particular, see fn. 40.

[34]To see how this works, first suppose that we have a selection function f and a base world w_1. Then, we can construct a *sequence* of possible worlds, $s = \langle w_1, ..., w_n \rangle$, corresponding to this selection function-world pair as follows. For any worlds w_i and w_j, let w_i come before w_j in the sequence just in case $f(\{w_i, w_j\}, w) = w_i$. Conversely, given s, we can construct a selection function as follows. Let $f(\{w_i, w_j\}, w) = w_i$ whenever w_i comes before w_j in s. The rest of f can then be derived from the constraints (i)–(iv) in the main text. Cf. Mandelkern (2018) and Khoo (2022).

The upshot is that, given a selection function and choice of base world w, there corresponds a *sequence* of possible worlds, which orders the worlds in $\mathcal{W}$ according to how similar they are to w. Conversely, for every sequence of worlds in $\mathcal{W}$, there corresponds a selection function-world pair, $\langle f, w\rangle$, such that w is the first world in the given sequence. What this means is that everything we could do before, using a selection function, we can now do using a sequence.

In particular, we can give a slightly different definition of Stalnaker's semantics (as I previously said).[35] To do so, let's first introduce some terminology. From here on out, let's say that a "factual" (i.e., non-conditional) proposition P is *true at a sequence*, s, just in case P is true at the *first* world in s. Then, let's say that a counterfactual $P > Q$ is true at s just in case Q is true at the first P-world in s.[36] Finally, let's say that $P > Q$ is true at a world, w, simpliciter just in case it's true at every (admissible) sequence whose first world is w. This, then, is our new definition of Stalnaker's semantics. (I'll return to the topic of admissibility in a moment.)

As an example, to make what I've just said bit more concrete, suppose that the set of worlds we're interested in is $\mathcal{W} = \{w_1, w_2, w_3\}$. Let $\mathcal{S}_\mathcal{W}$ be the set of all the sequences of worlds that we can generate from $\mathcal{W}$, namely:

$$\mathcal{S}_\mathcal{W} = \left\{ \begin{array}{l} \langle w_1, w_2, w_3\rangle, \langle w_1, w_3, w_2\rangle, \\ \langle w_2, w_1, w_3\rangle, \langle w_2, w_3, w_1\rangle, \\ \langle w_3, w_1, w_2\rangle, \langle w_3, w_2, w_1\rangle \end{array} \right\}.$$

Now suppose that P is a factual proposition true at the worlds w_1 and w_2, and Q is a factual proposition true at w_2 and w_3. Then, P is true at the first four sequences in $\mathcal{S}_\mathcal{W}$, as I've written it above; Q is true at the last four sequences; and $P > Q$ is true at the following sequences, since these are the only sequences whose first P-world is a Q-world: $\langle w_2, w_1, w_3\rangle$, $\langle w_2, w_3, w_1\rangle$, and $\langle w_3, w_2, w_1\rangle$. (Note also that, while $P > Q$ is true at the world w_2 simpliciter, it's neither nor true nor false at the world w_3, since there's one sequence beginning with w_3 whose first P-world is a Q-world, and there's one sequence beginning with w_3 whose first P-world is not a Q-world.)

Why, however, am I bothering to introduce this new formulation of Stalnaker's semantics, when the previous formulation seemed perfectly adequate? There are two key reasons. The first is simply that the constraints on selection functions that I gave above are all completely standard. Stalnaker himself assumes them, for example (1968). And so does nearly everyone who's worked with his semantics in the meantime. Thus, by appealing directly to sequences, rather than selection functions, we can forgo the need to keep mentioning the constraints (i)–(iv). They're built right into the sequence formulation of the semantics; they're not extra assumptions that we need to make.

The more important reason, however—as you're probably expecting—is that the sequence-based formulation of Stalnaker's semantics helps us to handle the issue of indeterminacy, which I mentioned at the end of the last section. To begin to see how, start by taking another look at the toy example, where $\mathcal{W} = \{w_1, w_2, w_3\}$. In that case, each world is consistent with two different similarity orderings. So, what this implies is that, even after we've pinned down truth-values for all the factual propositions at a world, we still haven't pinned down the truth-values for all the conditional propositions. Here's a picture, for illustration:

This is an idea worth dwelling on a bit. You'll notice that, in the sequence-based set-up I've just introduced, worlds no longer count as the most basic possibilities. Instead, sequences do. And we can think of worlds as sets of sequences, just as we could think, before, of propositions as sets of worlds. The idea is

[35] This isn't 100% accurate. As Matthew Mandelkern points out to me, the sequence formulation of Stalnaker's semantics requires a very mild strengthening of his background logic. However, this strengthening has no bearing on anything I'll say here, so it needn't concern us. See Mandelkern (forthcoming, §7.4) for further discussion.

[36] This definition only works for *simple* counterfactuals, i.e., those that don't have counterfactuals (or other modals) as antecedents or consequents. All the counterfactuals I'm interested in here, however, count as "simple" in this sense.

w_1		w_2		w_3	
w_1	w_1	w_2	w_2	w_3	w_3
w_2	w_3	w_1	w_3	w_1	w_2
w_3	w_2	w_3	w_1	w_2	w_1

Figure 2.2: Logical Space in the Toy Example

that, while all the "descriptive" facts are settled by the world, the conditional facts need not be. Instead, conditional facts depend for their truth on relations of similarity *between* worlds. Moreover, those relations are fixed by context; they needn't supervene on the descriptive facts that obtain at a world.

In the recent literature, this "fine-grained" view of a conditional's content has gained in popularity. Several authors have shown, for example, that it helps us to defuse puzzles arising from the interaction between our credences, on the one hand, and conditionals, on the other (see, e.g., Khoo and Santorio (2018), Goldstein and Santorio (2021), Khoo (2022), Schultheis (forthcoming), and Mandelkern (forthcoming)). Later in this section, I'll briefly mention one of those puzzles, and allude to how the fine-grained view helps to resolve it. But in the meantime, note that, since we're now working in a more fine-grained setting, we have to say how your credence function, cr, can be extended, so that it's defined over sequences, and not just over worlds.[37]

There are a number of ways we could make this extension, each with advantages and disadvantages.[38] But for simplicity, I'm here going to work with an idea from Goldstein and Santorio (2021) and Khoo (2022).[39] Specifically, I'll "lift" your credence function, cr, to a new credence function, pr, by means of a recursive procedure. This new function will then allow you to assign credences to arbitrary sets of sequences, and not just to sets of worlds.

To see how this works, let's start with some assumptions. First, let's suppose that the context set, $\mathcal{W}$, is just the set of your epistemically possible worlds. Then, let's assume that your credence function is *regular*, in the sense that, for every world $w \in \mathcal{W}$, your credence in w is such that $cr(w) > 0$. Neither of these assumptions is strictly essential for what I'm doing. But dropping them introduces additional complications which I'd rather not get into.

Now, in the simplest case, where all sequences of worlds count as admissible in the context, we can "lift" your credence function as in the following way (I'll give a slight refinement of this definition in a moment). First, let's write '$[w]$' for the set of sequences beginning with the world w, and '$[w_1, ..., w_k]$' for the set of sequences whose k-length initial segment consists of $w_1, ..., w_k$, in that order. Then, we define the credence function pr as:

(i) $pr([w]) = cr(w)$,

(ii) $pr([w_1, ..., w_k]) = pr([w_1, ..., w_{k-1}]) \cdot cr(w_k \mid \mathcal{W} - \{w_1, ..., w_{k-1}\})$.

Metaphorically, we can think of this as saying that your credence in a sequence $s = \langle w_1, ..., w_n \rangle$ is equal to your credence that you'd draw those worlds from an urn, in that order, and without replacement. For

[37] Do we also need to say how your value function, v, can be extended, so that it's defined over sequences? Not for present purposes, since I'll assume that outcomes are ordinary "factual" propositions. I explore this issue elsewhere, however. See McNamara (MS-b).

[38] See van Fraassen (1976) and Mandelkern (forthcoming) for proposals different to the one I'll make use of here.

[39] See also Khoo and Santorio (2018).

example, the credence that you assign to the sequence $\langle w_1, w_2, w_3\rangle$ from the toy example is just your credence that you'd draw w_1 first, multiplied by your credence that you'd draw w_2 second, having already drawn w_1, and so on. Since each sequence of worlds is supposed to correspond to an ordering of possible worlds according to how similar they are to a base world, it's easy to see why this lifting procedure is a sensible proposal.

It's also easy to see that pr preserves the credences that cr assigns to factual propositions. To quickly illustrate this anyway, however, consider again the toy example, where $\mathcal{W} = \{w_1, w_2, w_3\}$. Suppose that $cr(w_1) = cr(w_2) = cr(w_3) = 1/3$. Then, it follows that $cr(P) = 2/3$, since $P = \{w_1, w_2\}$. Now, by the definition of pr:

$$\begin{aligned} pr(\langle w_1, w_2, w_3\rangle) &= pr([w_1, w_2]) \cdot cr(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= pr([w_1]) \cdot cr(w_2 \mid \mathcal{W} - \{w_1\}) \cdot cr(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= cr(w_1) \cdot cr(w_2 \mid \mathcal{W} - \{w_1\}) \cdot cr(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= 1/3 \cdot 1/2 \cdot 1 \\ &= 1/6. \end{aligned}$$

Similar calculations show that $pr(\langle w_1, w_3, w_2\rangle) = pr(\langle w_2, w_1, w_3\rangle) = pr(\langle w_2, w_3, w_1\rangle) = 1/6$. And taking the sum of your credences in all of these sequences gives $pr(P) = 2/3$, as desired. (Note that your credence in a counterfactual $P > Q$ is also the sum of your credences in all of the sequences at which it's true. But in the present setting, this set of sequences need not always correspond to a set of worlds.)

Now, there's one last piece of the puzzle I need to put in place, before I can state my contextualist version of CDT precisely. Specifically, I need to say how things work out when not all of the possible similarity orderings are admissible in a context. After all, in §3.3 we heard that a similarity ordering is admissible only if it holds fixed the answers to a salient question. So, how are you supposed to assign credences to sets of sequences, given this constraint?

This turns out to be straightforward. For example, suppose we have a partition, $Z = \{Z_1, ..., Z_n\}$, corresponding to such a question. Then, a sequence of worlds $s = \langle w_1, ..., w_m\rangle$ corresponds to an admissible similarity ordering just in case all the worlds in s comes from a single cell Z_i. To illustrate this, let $\mathcal{W}$ again be the set $\{w_1, w_2, w_3\}$, and suppose the relevant partition is $Z = \{\{w_1, w_2\}, \{w_3\}\}$. Then, the admissible similarity orderings here are $\langle w_1, w_2\rangle$, $\langle w_2, w_1\rangle$, and $\langle w_3\rangle$, respectively, since these are the only orderings we can generate from the cells of the corresponding partition.

Given this constraint, we can give a slightly different definition of our lifting procedure:

(i) $pr([w]) = cr(w)$,

(ii*) $pr([w_1, ..., w_k]) = pr([w_1, ..., w_{k-1}]) \cdot cr(w_k \mid Z_i - \{w_1, ..., w_{k-1}\})$.

Here, Z_i is the partition cell to which w_k belongs. So the only difference between this definition of the lift of cr and our original definition is that, in this new case, once a base world has been chosen, we only consider worlds from the same partition-cell as that world. (In fact, our old definition and this new one agree, whenever the relevant partition is the trivial partition, $\mathcal{W}$.)

All this applies equally when we have more than one salient question in a context. For example, in cases where the similarity relations are constrained by the coarsest common refinement of two (or more) such questions, we can replace Z_i in the above definition with the cells from this coarsest common refinement. Similarly, when similarity relations are constrained by the finest common coarsening of some questions, we can replace Z_i with the cells from this latter partition. Like I said, the first of these cases won't play much of a role in what's to come. But the second one will be important in the next section, and I'll say more about it then.

For now, we're at last in a position to state the version of CDT that I've been working towards. In my view, when you're making a decision, you should choose an option that maximizes the following quantity, which I'll still refer to as *utility*:

$$U(A) = \sum_i pr(A > O_i) \cdot v(O_i). \tag{2.2}$$

This decision rule looks more-or-less identical to the original Stalnaker-Gibbard-Harper rule. The only difference is that, in the case of (2.2), the lifted credence function, pr, replaces the original credence function, cr. This means that the counterfactuals $A > O_i$ appealed to in this equation don't always have to correspond to a set of worlds. Instead, they can correspond to the set of all admissible similarity orderings at which that counterfactual is true. As we'll see later on, it's this change that helps us to get the right answer in cases involving indeterminacy.

Let me now close this section by making a few additional comments about the version of CDT I've just introduced.

First, you'll notice that, although I've been speaking throughout about *counterfactuals*, all of the worlds appealed to in my theory, as I've set it up here, are epistemically possible worlds. This, I think, is an important thing to point out, because some authors object to Stalnaker-Gibbard-Harper CDT on the grounds that it requires you to think about epistemically impossible worlds in certain deterministic cases. Kment (2023), for instance, criticizes CDT in this way, saying that epistemically impossible worlds are "irrelevant to a rational assessment of your options… Reflection on such worlds is a form of wishful thinking that has no place in rational choice" (p. 10). I'm not sure I agree with Kment about this for every decision problem (which worlds are relevant, after all, is a matter of context, in my view). But in any case, the objection has no force against my theory, since, as we'll see below, this theory gets the right answer in deterministic cases, and only appeals to epistemically possible worlds.[40]

Additionally, the formal framework I've set up here owes a lot to Khoo and Santorio (2018), Goldstein and Santorio (2021), Khoo (2022), and Mandelkern (forthcoming), all of whom use a similar framework for a very different purpose—namely, to prove *tenability results* for versions of *Stalnaker's thesis* (Stalnaker, 1970).[41] Recall that Stalnaker's thesis relates your credences in indicative conditionals to your conditional credences. Specifically, it says that, if you're rational, your credence in an indicative conditional $P > Q$ will match your conditional credence in Q given that P (assuming this is well-defined). Formally: $pr(P > Q) = pr(Q \mid P)$.[42] For a long time, it was thought that this thesis couldn't be true, owing to the famous triviality results of Lewis (1976) and others. But as the authors mentioned above have recently shown, versions of Stalnaker's thesis can hold (non-trivially) after all, provided all the sequences of possible

[40]Is it right to say that the conditionals in my theory are really then *counterfactuals*, rather than, say, indicative conditionals? You might be worried that they're the latter. However, I think it's still legitimate to call these conditionals 'counterfactuals' because the relations of similarity that are relevant to their assessment are those that hold fixed facts about causal connections, rather than, say, facts about (mere) epistemic connections. As Stalnaker (1975) and others have argued, the key difference between counterfactuals and indicative conditionals seems not to be anything to do with "counterfactuality" per se, but instead the fact that indicative conditionals are about epistemic possibilities, and counterfactuals are about causal or metaphysical possibilities.

[41]See also van Fraassen (1976) and Bacon (2015).

[42]I use the same symbol, '$>$', for both indicative conditionals and counterfactuals because Stalnaker's semantics is a *uniform* semantics. That is, it says that the truth-conditions for indicative conditionals and counterfactuals are one and the same, and all the differences between these conditionals come down to the salient similarity relations that we use to assess them. When all similarity relations are admissible in a context, however—and the context set consists just of epistemically possible worlds—Stalnaker's semantics says that these types of conditionals coincide. There's some evidence that this is indeed the case of natural language conditionals. For example, so-called "future-directed" counterfactuals often seem to say the same thing as corresponding indicative conditionals. Compare: 'If I were to flip the coin, it would land heads' and 'If I flip the coin, it will land heads'. Plausibly, the reason for this convergence is that we're using the same relations of similarity to assess these conditionals. See, e.g., Edgington (1995) for further discussion.

worlds count as admissible in a context. (If not all sequences are admissible, then the situation is more complicated. See, e.g., Khoo (2016, 2022) and Mandelkern (forthcoming).)

In the present setting, these tenability results turn out to have an interesting upshot. To see what it is, consider the following alternative to CDT's decision rule. Suppose that the quantity you should maximize when you're making a decision isn't U, but the following:

$$V(A) = \sum_i pr(O_i \mid A) \cdot v(O_i) \tag{2.3}$$

This quantity is sometimes called the *news value* of A. And it's the quantity that CDT's chief rival, *evidential decision theory* (EDT), tells you to maximize when you're choosing between your options.[43] Thus, what the tenability results I mentioned above imply is that, in any context in which all the similarity orderings are admissible, the version of CDT I've advocated for here will give the same recommendations as EDT. After all, in any case like that, $pr(A > O_i) = pr(O_i \mid A)$ for all O_i, and so $U(A) = V(A)$. This, I think, is a very interesting point of connection between my theory and a rival. And as we'll now see, it has important consequences for some of the decision problems we'll reconsider.

2.6 Cases Redux

With my version of Stalnaker-Gibbard-Harper CDT now in place, let's return to Ahmed's cases. In this section, I'll show that my theory gets the right answer in those cases. (After seeing this, it should also be obvious how my theory handles analogous deterministic cases, like those recently discussed by T. L. Williamson and Sandgren (forthcoming), Gallow (2022), Kment (2023), and others.) I'll also show that my theory gives the two-boxing recommendation in *Newcomb*. And I'll close the paper by considering a case that we haven't yet looked at.

2.6.1 *Betting on the Laws* Redux

Let's start with *Betting on the Laws*. In that case, you were offered a choice between two bets on the proposition L, namely: B_1, which pays \$1 if L is true, but pays nothing if L is false; and B_2, which pays nothing if L is true, but pays \$1 if L is false. Here, again, is the decision table:

	L	$\neg L$
B_1	\$1	\$0
B_2	\$0	\$1

[
Betting on the Laws]Table 2.2: *Betting on the Laws*

Now, given what I said in §3.3, it should be clear that the salient question here corresponds to $\{, \neg\}$. After all, the propositions in this partition are both causally independent of your choice; they're also consistent with each of your options; and any cell of this partition determines the amount of money you'll receive, once you've chosen a particular option. Thus, it follows that every admissible similarity ordering makes one of the following biconditionals true:

$$(B_1 > \$1) \equiv (B_2 > \$0),$$

$$(B_1 > \$0) \equiv (B_2 > \$1).$$

[43]EDT was first introduced by Richard R. C. Jeffrey (1965); see also R. C. Jeffrey (1983). Incidentally, Ahmed himself vigorously defends EDT over CDT, and sees his deterministic cases as giving us a reason to favor the former.

The sequences at which the first biconditional is true partition the proposition L. So your credences satisfy:

$$pr(B_1 > \$1) = pr(B_2 > \$0) = pr(L).$$

Similarly, the sequences at which the second biconditional is true partition the proposition $\neg L$. So your credences also satisfy:

$$pr(B_1 > \$0) = pr(B_2 > \$1) = pr(\neg L).$$

Given these equalities, we have:

$$\begin{aligned} U(B_1) &= pr(B_1 > \$1) \cdot 1 + pr(B_1 > \$0) \cdot 0 \\ &= pr(L) \cdot 1 + pr(\neg L) \cdot 0 \\ &= pr(L) \\ U(B_2) &= pr(B_2 > \$0) \cdot 0 + pr(B_2 > \$1) \cdot 1 \\ &= pr(L) \cdot 0 + pr(\neg L) \cdot 1 \\ &= pr(\neg L). \end{aligned}$$

Then, since $pr(L) \approx 1$ and $pr(\neg L) \approx 0$, it follows that $U(B_1)$
$approx 1$ and $U(B_2)$
$approx 0$. So my theory recommends B_1—the right answer.

Notice that, since the admissible sequences in this case partition the propositions L and $\neg L$, our calculations of utility simplified. Specifically, we ended up being able to calculate $U(B_1)$ and $U(B_2)$ directly in terms of your credences in the propositions L and $\neg L$. As it turns out, the same thing goes in any decision problem with similar features. That is, so long as there's just one salient partition in a decision problem, consisting of state-propositions whose truth-values are all causally independent of what you do, we can always calculate utility directly in terms of your credences in the states.

2.6.2 *Betting on the Past* Redux

Seeing what my theory says about *Betting on the Laws* was straightforward. But seeing what it says about *Betting on the Past* is a little trickier. After all, we saw in §3.3 that there isn't just one salient question here, but two. I'll repeat the relevant tables for convenience:

	H	$\neg H$
B_3	\$1	−\$10
B_4	\$10	−\$1

[
Betting on the Past]Table 2.3: *Betting on the Past*

	L	$H \wedge \neg L$	$\neg H \wedge \neg L$
B_3	\$1	\$1	−\$10
B_4	−\$1	\$10	−\$1

[
Betting on the Past]Table 2.5: *Betting on the Past*

Now, given that there's more than one salient question in this case, a natural first thought is that we should take the relevant similarity relations to be constrained, not by either of these questions alone, but by their conjunction—or more precisely, their coarsest common refinement. Unfortunately, however, this won't work, since the partition we end up with is one where your options are inconsistent with some of the cells, violating well-definedness. (See fn. 32 above, as well as J. M. Joyce (2016), Solomon (2021), Elga (2022), and Fusco (forthcoming) for further discussion.) So we need to try out something else.

Thus, consider the other thing I said in §3.3. There, I said that in certain contexts, it's more appropriate to think of similarity relations as being constrained by the *finest common coarsening* of questions, rather than by their coarsest common refinement. *Betting on the Past* shows, I think, why this is sometimes the

case. After all, the two questions here "compete" with one another, in the sense that holding fixed the answers to one question means you can't hold fixed the answers to the other—at least not when you're deliberating about what to do.

In a bit more detail: imagine you choose B_3 and then win you \$1. Then, you're almost certainly at a world where both H and L are true. But if that's so, then ask yourself: what would have happened if you had chosen the option B_4 instead? Here, it seems like you're pulled in two different directions. One salient question seems to imply that choosing B_4 would've won you \$10; but another seems to imply that choosing B_4 would've lost you a dollar. Thus, there seem to be different admissible precisifications of 'closest B_4-world in this case, which it makes it indeterminate what your choice of B_4 would've resulted in. In other words, different admissible precisifications say that different outcomes would've occurred, if you'd chosen otherwise than you actually did.

This, I think, is one of the things that makes *Betting on the Past* so interesting. In my view, the counterfactuals in this case admit of a significant amount of indeterminacy, in virtue of the two different questions in play. In order to capture this indeterminacy, we have to allow a whole range of similarity orderings to count as admissible. And the best way to do this, I believe, is to "merge" the salient questions, and think of similarity relations as being constrained by their finest common coarsening. This is the most plausible way I can see to allow, e.g., that there are non-actual B_4-worlds at which you win \$10, but also others at which you lose \$1, as in the above example. Similarly for the different precisifications of analogous counterfactuals.

But what *is* the finest common coarsening of the relevant questions in *Betting on the Past*? Well, since the only partition that's coarser than $\{H, \neg H\}$ is the trivial partition, $\mathcal{W}$, the only partition that's coarser than both of these questions is, again, the context set $\mathcal{W}$. Thus, on my analysis, *every sequence of worlds* is admissible in *Betting on the Past*. And this turns out to have an important upshot. Recall that in the previous section, I said it's implied by the tenability results for Stalnaker's thesis that, when all the sequences of possible worlds are admissible in a context, my theory gives the same recommendations as EDT. After all, in cases like that, we have $pr(A > O_i) = pr(O_i \mid A)$ for each O_i. So, given these equalities, we have:

$$\begin{aligned} U(B_3) &= pr(B_3 > \$1) \cdot 1 + pr(B_3 > -\$10) \cdot -10 \\ &= pr(\$1 \mid B_3) \cdot 1 + pr(-\$10 \mid B_3) \cdot -10 \end{aligned}$$

$$\begin{aligned} approx 1 \cdot 1 + 0 \cdot -10 & \\ &= 1 \\ U(B_4) &= pr(B_4 > \$10) \cdot 10 + pr(B_4 > -\$1) \cdot -1 \\ &= pr(\$10 \mid B_4) \cdot 10 + pr(-\$1 \mid B_4) \cdot -1 \end{aligned}$$

$$\begin{aligned} approx 0 \cdot 10 + 1 \cdot -1 & \\ &= -1. \end{aligned}$$

So, in the end, my theory says that $U(B_3)$
$approx 1$ and $U(B_4)$
$approx -1$. The theory thus recommends you choose B_3—the right answer.

2.6.3 *Newcomb* Redux

Contrast this with what my theory says about the *Newcomb* problem. In that case, the only salient question is *How much money is in the opaque box?*.[44] So it follows that every admissible similarity ordering is one which makes one of the following material biconditionals true:

$$(\textit{One-box} > \$1m) \equiv (\textit{Two-box} > \$1m + 1k),$$

$$(\textit{One-box} > \$0) \equiv (\textit{Two-box} > \$1k).$$

Given this, things play out much as they did in *Betting on the Laws*. That is, the sequences at which the first biconditional is true partition the proposition *Million*. So your credences satisfy:

$$pr(\textit{One-box} > \$1m) = pr(\textit{Two-box} > \$m + 1k) = pr(\textit{Million}).$$

Similarly, the sequences at which the second biconditional is true partition the proposition *No Million*. So your credences also satisfy:

$$pr(\textit{One-box} > \$0) = pr(\textit{Two-box} > \$1k) = pr(\textit{No million}).$$

This means that we can calculate utility straightforwardly using your credences in states:

$$\begin{aligned} U(\textit{One-box}) &= pr(\textit{Million}) \cdot 1m + pr(\textit{No Million}) \cdot 0 \\ &= pr(\textit{Million}) \cdot 1m \\ U(\textit{Two-box}) &= pr(\textit{Million}) \cdot (1m + 1k) + pr(\textit{No Million}) \cdot 1k \\ &= pr(\textit{Million}) \cdot 1m + 1k. \end{aligned}$$

The upshot is that $U(\textit{One-box}) < U(\textit{Two-box})$, no matter what your credences in *Million* and *No Million*. So my theory says that you should take both boxes, just as our original version of CDT did.

I think this demonstration shows that there's an important distinction between *Newcomb* and *Betting on the Past*. Specifically, in the former case, there's just one salient question raised by your decision situation; but in the latter case, there are two. Moreover, it's only if we focus on one of the questions that a causal dominance argument can be mounted in *Betting on the Past*. If we focus on the other question instead, then causal dominance reasoning doesn't apply. Thus, something Elga said earlier is plausibly mistaken. Recall his remark that:

> In a standard Newcomb problem there is a causal dominance argument for taking two boxes: 'The $1 million is either there or it is not, and you have no causal influence on whether it is. Either way (and no matter what else is true), taking two boxes gets you a better outcome than taking just one. So you should take two boxes.'... These conditions are satisfied in [*Betting on the Past*] just as much as they are in a standard Newcomb problem. (p. 207)

My sense, however, is that this isn't right. In *Newcomb*, *Two-box* causally dominates *One-box* on *every* admissible precisification of the counterfactuals. In *Betting on the Past*, in contrast, B_4 causally dominates B_3 only on some admissible precisifications. Thus, it isn't right to say that the relevant conditions "are satisfied in [*Betting on the Past*] just as much as they are in a standard Newcomb problem". So, in my view, causal decision theorists have a good reason to resist Elga's conclusion.

[44] Zach Barnett asks me why we can't think of the partition {*The predictor is accurate, The Predictor is inaccurate*} as corresponding to a salient question in *Newcomb*. One answer is that we can—although it's not one that can constrain the similarity relations for *standard* counterfactuals. The reason is that counterfactuals whose similarity relations are constrained by this partition are plausibly true only on a backtracking interpretation. And in §2.2, I set backtracking counterfactuals aside. Along the same lines, it's plausible that the the propositions in this partition aren't really causally independent of your options. The reason, as J. M. Joyce (2018) points out, is that it's an important feature of any genuine *Newcomb*-type problem that you believe you have the power to make the predictor's prediction wrong—even if you're certain you won't actually do this.

2.6.4 A New Case

I'm going to wrap up now by considering a new case. I'm adapting it from one recently discussed by Hedden (2023). And similar cases have been considered by Solomon (2021), Gallow (2022), and Fusco (forthcoming), among others. The case is also a bit like the one we considered right at the outset:

> *Betting on the Past and the Laws.* You have to choose between two bets on a proposition D. First, there's B_5, which pays \$1 if D is true, but loses \$10 if D is false. Second, there's B_6, which pays \$10 if D is true, and loses only \$1 if D is false. D is the proposition that the past state of the world, together with the laws of nature, determines that you accept B_5. And because you're certain of determinism, you're certain that the negation of this proposition determines that you accept B_6.

	D	$\neg D$
B_5	\$1	$-$\$10
B_6	\$10	$-$\$1

Table 2.6: *Betting on the Past and the Laws*

On its face, this case looks a little bit like *Betting on the Past*.[45] However, there's an important difference between that case and this new one. Unlike in *Betting on the Past*, the proposition D here is *inconsistent* with your choosing B_6, while $\neg D$ is inconsistent with your choosing B_5. So, the only way you won't lose a dollar by choosing B_6 is if a contradiction is true.

Hedden claims that various formulations of CDT tell you to choose B_6 in *Betting on the Past and the Laws*. After all, he says, the proposition D

> is a proposition about the [history] of the universe and the laws of nature. The [history] of the universe and the laws of nature are both beyond your causal control; you cannot cause either one to be different. Therefore, no matter how things beyond your causal control might be (i.e. no matter whether D or $\neg D$ is true), B_6 yields a strictly better outcome than B_5. (2023, p. 743, notation adapted)

But more than any case we've seen, this is clearly absurd.

Now, I'm not convinced that CDT, in any formulation, tells you to choose B_6, contrary to what Hedden says. (Although I'll concede that some versions of the theory may simply fall silent—think, for instance, about the "dependency hypothesis" versions of CDT due to Lewis and Skyrms.) However, setting aside my worries about Hedden's applications of CDT, I want to show that my theory gets the right answer in this case anyway.

The reasoning here is straightforward. Once more, recall from §3.3 that I said that, in any decision situation, the salient question must be one for which each cell of the corresponding partition is consistent with each your options. Thus, given this way of thinking about salient questions, it should be clear that there's only one such "question" that can be said to be raised in this situation. This is just the trivial question, $\mathcal{W}$ itself. So your decision problem looks as follows:

[45] Some authors—Hedden among them, but also Stalnaker (MS)—say that this case just *is Betting on the Past*. But I disagree. As Ahmed (2014b) is careful to say, for example, the proposition H in *Betting on the Past* is not meant to be a "cheesy" proposition, of the form 'The past and the laws, whatever they are, determine that you accept B_3'. Rather, H is quite specific in what it describes—namely, the intrinsic state of the world at some particular time in the distant past. This is the reason my treatment of *Betting on the Past* got quite complicated. That said, even though I think the case given here is distinct from *Betting on the Past*, my treatment of it is broadly similar to Stalnaker's.

	$\mathcal{W}$
B_5	\$1
B_6	$-$\$1

Table 2.7: *Betting on the Past and the Laws*

Then, since there's just one outcome consistent with B_5, and one outcome consistent with B_6; and since the first outcome is clearly better than the second; my theory tells you to choose B_5.[46] The upshot is that, even if Hedden is correct to say that other versions of CDT get this case wrong, my theory gets it right. It tells you to choose the option that has the best *possible* outcome.

[46]In a bit more detail: since the salient question here is just the trivial question, $\mathcal{W}$, all sequences of possible worlds count as admissible in the context. So, by the tenability results for Stalnaker's thesis, it follows that we can calculate utility using your conditional credences, rather than your credences in counterfactuals. Then, since $pr(\$1 \mid B_5) = 1$ and $pr(-\$1 \mid B_6) = 1$, it follows that $U(B_5) = 1$ and $U(B_6) = -1$. Thus, my theory tells you to choose B_5.

Chapter 3

Desire-as-Belief in Context

3.1 Introduction

How much should you desire that a proposition, A, be true? A plausible answer is: to the same extent you believe it would be *good* if A were true. This simple and compelling idea has a venerable history in philosophy.[1] David Lewis (1988a, 1996) called it the *Desire-as-Belief (DAB) thesis*.

Unfortunately, while the DAB thesis seems plausible on a first pass, Lewis also showed that it faces a serious difficulty. Given only mild assumptions, the thesis turns out to clash with a widely accepted formulation of decision theory—namely, the formulation due to Richard R. C. Jeffrey (1965, 1983). As Lewis says, Jeffrey's "Theory is an intuitively convincing and well worked-out formal theory of belief, desire, and what it means to serve our desires according to our beliefs. It is of course idealized, but surely it is fundamentally right. [So, if the DAB thesis] collides with Decision Theory, it is the Desire-as-Belief Thesis that must go" (1988a, p. 325).

This result seems like bad news for those of us who are attracted to the DAB thesis. However, in this paper, I want to argue that Lewis's result can be resisted. To do this, I'm going to draw a connection—one that's been drawn in the literature several times before, not least by Lewis himself—between the DAB argument, on the one hand, and Lewis's own *triviality results* for *Stalnaker's thesis*, on the other. Roughly, Stalnaker's thesis says that your credence in an indicative conditional 'If A, then C' should match your conditional credence in C, given that A. The triviality results then purport to show that this thesis can't, in general, be true. Given only mild assumptions, analogous to those we made in the case of the DAB argument, Stalnaker's thesis can hold only in trivial cases.[2]

Like the DAB thesis, however, Stalnaker's thesis is prima facie plausible. And this has led some philosophers to question whether Lewis's "mild assumptions" are really so mild. One of those assumptions in particular is problematic—namely, the assumption that indicative conditional sentences express the same proposition *in every context*. Opposed to this, a long tradition in philosophy (and semantics) claims that indicative conditional sentences are *context-sensitive*. That is, which proposition they express depends on the context in which they're uttered.[3] As it turns out, this *contextualist* view about indicative conditionals

[1] Versions of this idea have been defended by, e.g., Plato, Aristotle, Spinoza, Leibniz, and arguably even Hume (see Gregory, 2017, §1.7 and references therein). Furthermore, precisifications of this idea have been defended more recently by Price (1989), Broome (1991), Oddie (1994), Byrne and Hájek (1997), Hájek and Pettit (2004), Bradley and List (2009), Collins (2015), Bradley and Stefánsson (2016), and Gregory (2017), among others.

[2] See Stalnaker (1970) for the original statement of Stalnaker's thesis, and Lewis (1976) for the original triviality results. Lewis's results were later strengthened by, e.g., Hájek (1989, 2012), Hájek and Hall (1994), Bradley (2000), Fitelson (2015), and Goldstein and Santorio (2021).

[3] It's not universally agreed that indicative conditionals express propositions. In particular, so-called *expressivists*, like Adams

can be used to block Lewis's argument against Stalnaker's thesis. And better still, several philosophers have shown that, by appealing to contextualism, we can actually *prove* that Stalnaker's thesis holds for an important class of indicative conditionals.[4]

Given the parallels between Lewis's triviality results for Stalnaker's thesis and his argument against the DAB thesis, it's not surprising that philosophers have sometimes responded to the latter in ways that parallel responses to the former. However, it *is* surprising that a contextualist response to the DAB argument, analogous to the contextualist response to the triviality results, hasn't been explored yet in the literature. My aim here is to fill that gap. In §4.2 I'll start by reviewing Lewis's argument against the DAB thesis and his triviality results for Stalnaker's thesis. In §3.3 I'll make a case for a contextualist version of the DAB thesis, arguing that this version is independently plausible. I'll also show that we can use the contextualist version of the DAB thesis to block Lewis's argument. And in §3.4 I'll sketch a tenability result for my version of the DAB thesis, showing that there are non-trivial cases in which it holds. (A proof of this result can be found in the Appendix.) In §3.5 I'll compare my response to Lewis to a couple of other responses that have appeared in the philosophical literature. And in §3.6 I'll wrap things up by discussing some residual cases in which there remains a tension between Jeffrey's decision theory and the DAB thesis. While these cases show that DAB can't hold in every context, it can still hold in many. And this is a far cry from Lewis's claim that it can't hold, non-trivially, at all.

3.2 Background

3.2.1 Lewis on Desire-as-Belief

Let's begin with some assumptions. In what follows, I'll assume that your credences (or 'degrees of belief') at any time can be represented by a probability function, p, which I'll call your *credence function*. I'll assume that your desires at any time can be represented by a real-valued function, v, called your *subjective value function*.[5] I'll assume that p and v are both defined over a space of possible worlds, $\mathcal{W}$. And, for reasons that will become clear later on, I'll assume that $\mathcal{W}$ is finite, and also that p is *regular*, in the sense that $p(w) > 0$ for each world w in $\mathcal{W}$.[6]

Let A now be any proposition (where a *proposition*, for the moment, is a set of possible worlds, i.e., a subset of $\mathcal{W}$). Let A° (pronounced "A-halo") be the proposition expressed by 'A's truth would be good'. Then, the DAB thesis, stated formally, says the following relationship holds between v and p, if you're rational:

$$v(A) = p(A^\circ). \tag{DAB}$$

In words: your degree of desire for A's truth should match your degree of belief that A's truth would be good.

(1975), Gibbard (1981), Edgington (1995), Bennett (2003), and Moss (2015, 2018), think that indicative conditional sentences are expressions of your conditional credences. Unfortunately, I don't have space to engage with this view here. So instead, I'll simply assume that indicative conditionals express propositions in what follows. That said, we'll see in §3.4 that what I mean by 'proposition' is a bit non-standard—I won't be taking propositions to be sets of possible worlds from that point on.

[4]For results of this kind, see below, as well as, e.g., Bacon (2015), Khoo and Santorio (2018), Khoo (2022), and Mandelkern (forthcoming). There are also versions of these "tenability results" that don't rely on contextualism. See, e.g., McGee (1989), Stalnaker and Jeffrey (1994), Kaufmann (2004), Bradley (2012), and Goldstein and Santorio (2021). The first tenability result for Stalnaker's thesis was proved by van Fraassen (1976). Van Fraassen himself isn't explicit about how a contextualist view fits with his result. But it's natural to view his tenability proof through a contexualist lens (albeit quite an extreme one—see fn. **??** below as well as Bacon (2015) for further discussion).

[5]In this paper, I'm going to use the terms 'subjective values' and 'desires' interchangeably. Similarly for 'value' and 'desire' taken as verbs. This usage arguably elides some important distinctions. But those distinctions aren't important for our purposes here. Besides, Lewis himself often use the terms 'value' and 'desire' interchangeably in his discussions of the DAB thesis (see, e.g., Lewis, 1988a, p. 326). So, in a sense, I'm following his lead.

[6]I write '$p(w)$' instead of '$p(\{w\})$' to ease notion. Similarly for '$v(w)$'.

There are a couple of things to note about this statement of the DAB thesis. The first is that, as I'm construing it, DAB is a *normative* thesis, rather than a descriptive one. This is a bit different to how Lewis himself thought of the thesis. According to him, DAB is an "anti-Humean" thesis, which is supposed to hold as a matter of necessity. As Bradley and Stefánsson (2016) point out, however, this interpretation faces challenges: "if DAB were a psychological claim about ordinary people, then we wouldn't need a philosophical or decision-theoretic argument to examine its plausibility: citing ordinary psychological experience (with all its confused desires and so on) would then suffice to refute the thesis" (p. 694). Since I agree with Bradley and Stefánsson about this, I'll only consider DAB on its normative interpretation here. (Although the arguments I give in §§3.3–3.4 are compatible with Lewis's anti-Humean interpretation.)[7]

The other thing to note is that my statement of the DAB thesis is simplified in an important way. Specifically, in taking A° to be the proposition expressed by 'A's truth would be good', I've implicitly assumed that goodness doesn't come in degrees. This, of course, is implausible. But it's also merely a simplifying assumption. In §3.4, and in the Appendix, I consider a more realistic version of the DAB thesis, which allows for degrees of goodness. I focus on the simplified version here—following Lewis—just to streamline the discussion.

Now, provided these points about the DAB thesis are acknowledged, I think this thesis is very plausible. This is especially the case if you're a *realist* about value—i.e., you believe that at least some values are objective and mind-independent. To see why this is, consider some remarks from Oddie (2001):[8]

> For the realist, it is easy to provide both conceptual space, and motivation, for [the DAB thesis]. The realist takes there to be genuine truths about what is valuable. Consequently the realist thinks there are two regulative ideals which should constrain our valuing. First, one's beliefs about the good should be true. Second, one should desire things in proportion to their value… [I]t is pretty obvious that these two ideals together enjoin [the DAB thesis]. (p. 110)

As I said before, however, Lewis argues that DAB is in tension with decision theory. And this casts doubt on its credibility (and some believe, on realism more generally).[9] To see how his argument works, let's first note that in Jeffrey's decision theory, your degree of desire for A's truth is defined to be the following quantity (R. C. Jeffrey, 1965, 1983):

$$v(A) = \sum_w p(w \mid A) \cdot v(w). \qquad \text{(Subjective Value)}$$

Here, $p(w \mid A)$ is your conditional credence that w is actual, given that A is true, and $v(w)$ is your degree of desire that w be actual.[10] Very roughly, then, this definition says that A's subjective value is a weighted average of the values of the different ways in which A can be true. (This definition will be especially important in §3.2.2.)[11]

[7] See Bradley and Stefánsson (2016) for further discussion of why DAB is best interpreted as a normative thesis. See also Gregory (2017) for a recent defense of the alternative anti-Humean view.

[8] Note that Oddie uses the term 'Harmony' in place of 'the DAB thesis'. However, it's clear from his discussion that he thinks the notion of Harmony is equivalent to the DAB thesis, or at least that the DAB thesis entails Harmony. Note also that Bradley and Stefánsson (2016) argue that DAB needn't be interpreted as a realist thesis. But again—and as Oddie makes clear—the realist perhaps has the easiest time motivating the DAB thesis.

[9] As Oddie (1994) says, if Lewis's argument "is sound then it gets as close to being a reductio of realism about value as any argument could be, a rigorous proof of Wittgenstein's Tractarian claim that 'it is impossible for there to be propositions of ethics'"(p. 452).

[10] We define conditional credences using the standard ratio formula: $p(C \mid A) = p(A \wedge C)/p(A)$, for any propositions A and C. This is defined only if $p(A) > 0$. But for simplicity, I'll assume that all conditional credences under discussion here are defined.

[11] Note that Jeffrey's equation in the main text is closely related to the decision rule of so-called *evidential decision theory* (EDT). This might lead you to believe that Lewis's argument doesn't affect EDT's main rival, *causal decision theory* (CDT). As

In addition to adopting this definition of subjective value, Lewis makes two background assumptions before giving his DAB argument. The first is that you update your credences by standard *Conditionalization*. To spell this out: let p_A be your credence function after learning A. Then conditionalization says, for any proposition:

$$p_A(-) = p(- \mid A) \tag{Conditionalization}$$

That is, your credence in any proposition after learning A is equal to your prior credence in that proposition, conditional on A.

The other assumption Lewis makes is one he calls *Invariance*. This is the assumption that your subjective value for A doesn't change when you learn that A. In symbols, where v_A is your subjective value function after learning A:

$$v_A(A) = v(A). \tag{Invariance}$$

Now, this assumption has met with some resistance in the literature (see, e.g., Weintraub, 2007; Bradley and List, 2009; Stefánsson, 2014; and Bradley and Stefánsson, 2016). However, I think it's a natural assumption for realists about value to make. Lewis himself defends it, for example, by pointing out that Invariance holds for a proposition if it holds for all the maximally specific subcases of that proposition—viz., the possible worlds which are its members. But then, if a subcase

> were maximally specific merely in all 'factual' aspects... it would be no surprise if a change in belief changed our minds about how good it would be [if the subcase were true]... But the subcase was supposed to be maximally specific in *all* relevant aspects... [In other words] The subcase has a maximally specific hypothesis about what would be good built right into it. So in assigning it a value, we do not need to consult our opinions about what is good. We just follow the built-in hypothesis. (Lewis, 1988a, p. 332, emphasis added)

In §3.6, I'll have more to say about Invariance. There, we'll see that there's a sense in which it's *implied* by Jeffrey's theory, and so isn't really an extra assumption that Lewis has to make at all. However, getting clear on why that is requires some additional ideas, which we've not yet introduced. So, in the meantime, let's just see how Lewis's argument works once we've taken Invariance on board. Consider:

$$\begin{aligned} p(A^\circ) &= v(A) && \text{(DAB)} \\ &= v_A(A) && \text{(Invariance)} \\ &= p_A(A^\circ) && \text{(DAB)} \\ &= p(A^\circ \mid A) && \text{(Conditionalization)} \end{aligned}$$

Spelled out, this says that, in the presence of Invariance and Conditionalization, the DAB thesis demands that your credence in A° be *probabilistically independent* of your credence in A. But unless the range of credences you assign to propositions is extraordinarily impoverished, this can't always be the case.[12] There are many propositions you could learn which would cause DAB to go from satisfied to unsatisfied.[13]

many philosophers have pointed out, however, while causal decision theorists deny that Jeffrey's equation is a good measures of *choiceworthiness* for actions, it's plausibly still the correct measure of rational subjective *value*. In a sense, then, this equation—understood as a measure of subjective value rather than choiceworthiness—is common to both EDT and CDT.

[12] As McGee (1989) points out, if your credence function satisfies this constraint, then it assigns one of only four values (at most) to every proposition. But no such credence function can plausibly represent the full range of credences you assign to propositions. A related result of Costa et al. (1995) shows that, provided there are three pairwise inconsistent propositions, each of which you assign non-zero credence, if you satisfy the DAB thesis, then you must view with *indifference* any proposition whose probability is greater than 0. That is, $v(A) = v(\neg A)$ for any proposition A such that $p(A) > 0$.

[13] Here's a classic example. Suppose that $0 < p(A), p(A^\circ) < 1$, and $p(A \vee A^\circ) < 1$. Suppose also that $p(A^\circ) = p(A^\circ \mid A)$. Now imagine that you learn the disjunction $A \vee A^\circ$. Then, since you update your credences by Conditionalization, all of your credence in $\neg(A \vee A^\circ)$ will be redistributed to $A \vee A^\circ$, in such a way that $p(A^\circ \mid A) = p_{A \vee A^\circ}(A^\circ \mid A)$. (This follows from what's called the *rigidity* property of Conditionalization.) However, after learning $A \vee A^\circ$, your credence in A° will increase. So, $p_{A \vee A^\circ}(A^\circ) > p_{A \vee A^\circ}(A^\circ \mid A)$. Thus, after learning $A \vee A^\circ$, the indepedence between A and A° no longer holds.

Thus, the DAB thesis, it seems, can't be true in general.

3.2.2 Stalnaker's Thesis and Triviality

A number of philosophers have noted the close connection between the foregoing argument against the DAB thesis, and Lewis's own triviality results for Stalnaker's thesis.[14] To illuminate this connection, then, let me first state Stalnaker's thesis precisely. For starters, let $>$ be an operator which takes propositions A and C, and returns the proposition $A > C$, expressed by the indicative conditional sentence 'If A, then C'. Then, Stalnaker's thesis says:

$$p(A > C) = p(C \mid A). \qquad \text{(Stalnaker's Thesis)}$$

Informally: your credence in the proposition expressed by 'If A, then C' should match your conditional credence in the consequent, given the antecedent.

Versions of this thesis have been widely defended in the literature.[15] To see why, consider the following example. Suppose I'm about to roll a fair, six-sided die, when I say to you:

(1) If this die doesn't land on 1, then it will land on 2.

How confident should you be of the truth of this sentence? Here, most people say that your credence should $1/5$. And notice that, if you give equal credence to each possible outcome of the die roll, then this is just your conditional credence in 2, given that the die doesn't land on 1—which is what Stalnaker's thesis requires.

Still, while Stalnaker's thesis seems plausible on a first pass, Lewis's triviality results show that it can't be true in general, given only the assumption that you update your credences by Conditionalization (and also the anti-contextualist assumption that I mentioned in §3.1—but ignore that for now). To see why, consider the following derivation:

$$\begin{aligned}
p(C \mid A) &= p(A > C) && \text{(Stalnaker's Thesis)} \\
&= p(A > C \mid C) \cdot p(C) + p(A > C \mid \neg C) \cdot p(\neg C) && \text{(Law of Total Probability)} \\
&= p_C(A > C) \cdot p(C) + p_{\neg C}(A > C) \cdot p(\neg C) && \text{(Conditionalization)} \\
&= p_C(C \mid A) \cdot p(C) + p_{\neg C}(C \mid A) \cdot p(\neg C) && \text{(Stalnaker's Thesis)} \\
&= p(C \mid A \wedge C) \cdot p(C) + p(C \mid A \wedge \neg C) \cdot p(\neg C) && \text{(Conditionalization)} \\
&= 1 \cdot p(C) + 0 \cdot p(\neg C) && \text{(Probability Theory)} \\
&= p(C)
\end{aligned}$$

Thus, in parallel to what we proved in the last subsection, we just proved that if you satisfy Stalnaker's thesis at various times, then A and C must be independent in your credence function. But again, provided the range of credences you assign to propositions isn't totally impoverished, this can't always be the case. There are many propositions you could learn which would cause Stalnaker's thesis to go from satisfied to unsatisfied.[16]

There is, then, a clear parallel between Lewis's argument against the DAB thesis, and his triviality results for Stalnaker's thesis—the upshot of the two arguments turns out to be more-or-less the same. Still,

[14] See, e.g., Lewis (1988a, p. 329), and also Broome (1991), Oddie (1994), Hájek and Pettit (2004), Bradley and Stefánsson (2016), and especially Hájek (2015).

[15] Among many others, see Stalnaker (1970, 1984), McGee (1989), Bacon (2015), Khoo and Santorio (2018), Goldstein and Santorio (2021), Khoo (2022), Schultheis (2023), and Mandelkern (forthcoming).

[16] Consider, for instance, the case in which you learn the disjunction $A \vee (A > C)$. Then we can use an argument similar to the one given in fn. 13 to show that, even if you satisfy Stalnaker's thesis before learning this proposition, you won't satisfy it after.

we can make that connection even closer. To see how, start by noting that Lewis himself says there's a more conspicuous way in which we can construe the proposition A°. We first let G be the proposition consisting of all and only "the objectively desirable [worlds]—for short, the *good* ones" (Lewis, 1996, p. 307). We then let $A > G$ be the proposition expressed by the sentence 'If A, then G'—or, if you like: 'If A, then things are good'.[17] Now, given this proposition, we can re-state the DAB thesis in a slightly different way (Lewis, 1988a, p. 329):

$$v(A) = p(A > G). \tag{Conditional DAB}$$

This says that your desire for A should match your credence in the proposition expressed by the sentence 'If A, then G'.

Suppose, however, that your subjective values are scaled appropriately, so that, for every world $w \in G$, we have $v(w) = 1$.[18] Suppose further that, for every world $w \in \neg G$, we have $v(w) = 0$. Then, by Jeffrey's definition of rational subjective value, it follows that $v(G) = 1$ and $v(\neg G) = 0$. Moreover:

$$\begin{aligned} v(A) &= p(G \mid A) \cdot v(G) + p(\neg G \mid A) \cdot v(\neg G) \\ &= p(G \mid A) \cdot 1 + p(\neg G \mid A) \cdot 0 \\ &= p(G \mid A). \end{aligned}$$

Now it's easy to see the problem. Jeffrey's definition of subjective value implies that $v(A) = p(G \mid A)$. But the DAB thesis says that $v(A) = p(A > G)$. The two together imply that $p(A > G) = p(G \mid A)$. But this is just an instance of Stalnaker's thesis. And Lewis's triviality results seem to show that Stalnaker's thesis can't be true.

3.3 Contextualism

Faced with these arguments against the DAB thesis, some philosophers have concluded that this thesis should be abandoned (Costa et al., 1995; Ahmed and Spencer, 2020; Ahmed, 2021). However, others have tried to resist Lewis's arguments in various ways. For instance, some have claimed that we should reject the assumptions that underpin those arguments—either Invariance (Weintraub, 2007; Bradley and List, 2009; Stefánsson, 2014; Bradley and Stefánsson, 2016) or Conditionalization (Oddie, 1994). Others have said that we should replace Jeffrey's definition of rational subjective value with something else (Byrne and Hájek, 1997; Collins, 2015). And others still have introduced alternative versions of the DAB thesis, intended to circumvent Lewis's argument (Price, 1989; Broome, 1991; Hájek and Pettit, 2004). I'll have more to say about a couple of these responses later on. But for now, just note that some of them bear similarities to responses to the triviality results. For instance, Oddie's (1994) response, which rejects Conditionalization, is similar to a response to the triviality results given by McGee (1989) and Goldstein and Santorio (2021),

[17] You may have noticed here that, while I'm taking $A > G$ to be an *indicative* conditional, my original, informal statement of the DAB thesis used *counterfactual* language. Ultimately, I think that a fully general version of the DAB thesis should apply to counterfactual conditionals, and I discuss this issue elsewhere (McNamara, MS-a; see also Bradley and Stefánsson, 2017). However, for present purposes, notice that, when A is epistemically possible, in the sense that $p(A) > 0$, the indicative conditional 'If A, then C' often seems to coincide with the corresponding counterfactual 'If A, would C'. For example: 'If I flip this coin, it will land heads' and 'If I were to flip this coin, it would land heads'. Furthermore, since Jeffrey's theory says that $v(A) = p(G \mid A)$, and since the right-hand side of this equation is defined only when $p(A) > 0$, it's natural to interpret Jeffrey's decision theory as applying only to cases where A is epistemically possible. Thus, taking $A > G$ to be an indicative conditionals seems appropriate in this setting. Note also that the semantics for conditionals I give in §3.4—which is based on Stalnaker's (1968) semantics—is a *uniform* semantics, in the sense that it gives the same truth-conditions to indicatives and counterfactuals. Thus, in that case, we can interpret $A > G$ either as indicative conditional or a counterfactual (assuming, again, that A is epistemically possible).

[18] In general, the numbers we use to represent subjective values are unique only up to positive affine transformation. So, we can always scale your values in such a way that the condition stated in the main text holds. See Hájek and Pettit (2004) for further discussion.

which also rejects Conditionalization.[19] Analogously, Hájek and Pettit (2004) say that we should resist Lewis's DAB argument by embracing an "indexical" version of the DAB thesis. And this is similar to the "indexicalist" response to the triviality results given by van Fraassen (1976) and Stalnaker and Jeffrey (1994) (something that Hájek and Pettit acknowledge—see §3.5).

The fact, however, that responses to Lewis's DAB argument often parallel responses to his triviality results makes it surprising that a *contextualist* response to the former hasn't been given yet in the literature. After all, to many of us, the contextualist response to the triviality results seems like the strongest such response available. It's been given by, e.g., Bacon (2015), Khoo and Mandelkern (2018), Schultheis (2023), and Mandelkern (forthcoming), among others. And it's worth seeing briefly how it works.

Thus, to start off, recall my earlier statement of Stalnaker's thesis. I said: your credence in the proposition expressed by 'If A, then C' should match your conditional credence in the consequent, given the antecedent. Now, with that in mind, consider a few lines from Lewis's proof:

$$\begin{aligned}
p(C \mid A) &= p(A > C) && \text{(Stalnaker's Thesis)} \\
&= p(A > C \mid C) \cdot p(C) + p(A > C \mid \neg C) \cdot p(\neg C) && \text{(Law of Total Probability)} \\
&= p_C(A > C) \cdot cr(C) + p_{\neg C}(A > C) \cdot p(\neg C) && \text{(Conditionalization)} \\
&= p_C(C \mid A) \cdot cr(C) + p_{\neg C}(C \mid A) \cdot p(\neg C) && \text{(Stalnaker's Thesis)} \\
&\vdots
\end{aligned}$$

Here, you'll notice that there's an implicit assumption that $A > C$ is the proposition expressed by 'If A, then C' *both before and after* you've learned one of C or $\neg C$. Without that assumption, the proof doesn't go through, since we can't infer the fourth line from the third. (After all, it's only the proposition expressed by 'If A, then C' to which Stalnaker's thesis applies.) However, contextualists about indicative conditionals won't grant this assumption. Instead, they'll say that 'If A, then C' can express different propositions before and after you've learned C or $\neg C$, in just the same way that 'Today is Tuesday' expresses different propositions depending on which day of the week it's said. (If it's uttered on a Tuesday, for example, it expresses a true proposition, while if it's uttered on a Wednesday, it expresses a false proposition.) Thus, for contextualists, Lewis's argument doesn't establish its conclusion—we can block his proof at the transition from the third to the fourth step.

Of course, the plausibility of this response hangs on there being good reasons to think that contextualism about indicative conditionals is plausible in the first place. So let me just outline one reason to think that this is so. One prominent version of contextualism—the kind that's of most interest here—says that the proposition expressed by a sentence 'If A, then C' depends on a body of evidence that's salient in the context.[20] And a natural source of motivation for this view comes from so-called *stand-off cases*, like this one, from Gibbard (1981):

> Sly Pete and Mr. Stone are playing poker on a Mississippi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good, and signals its content to Pete. My henchman Jack sees both hands and sees that Pete's hand is rather low, so that Stone's is the winning hand. At this point, the room is cleared. A few minutes later, Zack slips me a note which says, 'If Pete called, he won,' and Jack slips me a note which says 'If Pete called, he lost.' I know that both notes come from my trusted henchmen but do not know which of them sent which note. I conclude Pete folded. (p. 231)

[19] See also Boylan (MS).

[20] See, e.g., Stalnaker (1975), Bacon (2015), Boylan and Schultheis (2021), or Schultheis (2023) for more on how the proposition expressed by 'If A, then C' depends on a *body of evidence* that's salient in the context.

Here, contextualists think that Zack and Jack *both* speak truly when they say 'If Pete called, he won' and 'If Pete called, he lost', respectively. The reason is that different bodies of evidence are available to the speakers in the two different contexts. In Zack's case, his evidence includes the proposition that Pete knows the hand that Mr. Stone has been dealt (and nothing else of relevance). So, in his context, the proposition expressed by 'If Pete called, he won' seems true. Conversely, Jack's evidence includes the proposition that Pete's hand is lower than Mr. Stone's. And that's why it sounds true for him to say 'If Pete called, he lost'. Of course, there's no context in which 'If Pete called, he won' and 'If Pete called, he lost' are both true (at least not given a plausible principle of conditional non-contradiction, which I'll assume). And this is why Gibbard can conclude that, in fact, Pete folded.

I don't have space to give a complete defense of this version of contextualism here. But I hope the case above at least makes that view seem plausible.[21] Going forward, I'll mostly assume this form of contextualism about indicative conditionals. What I want to do now is show how a similar view about sentences like 'A's truth would be good' can be used to block Lewis's argument against the DAB thesis. Later on, we'll see that this view can even be used to show that the DAB thesis is *tenable*, in the sense that there are non-trivial cases in which it holds.

3.3.1 Contextualism and the DAB Thesis

If sentences like 'A's truth would be good' are context-sensitive sentences, then it's easy to adapt the contextualist response to the triviality results above to the case of the DAB argument. To see how, first recall my initial statement of the DAB thesis. I said that your subjective value for A should match your credence in the proposition expressed by the sentence 'A's truth would be good'. Now, with that in mind, let's take another look at Lewis's proof from §4.2:

$$\begin{aligned} p(A^\circ) &= v(A) && \text{(DAB)} \\ &= v_A(A) && \text{(Invariance)} \\ &= p_A(A^\circ) && \text{(DAB)} \\ &= p(A^\circ \mid A) && \text{(Conditionalization)} \end{aligned}$$

In parallel to what we saw in the last subsection, there's an implicit assumption here that A° is the proposition expressed by 'A's truth would be good' both before and after you've learned the truth of A. Without that assumption, the third line of the proof doesn't follow, since it's only the proposition expressed by 'A's truth would be good' to which the DAB thesis applies. However, if this sentence is context-sensitive, then this step in the proof isn't one we should accept. Instead, 'A's truth would be good' might express different propositions in the two different contexts. So again, it looks like the upshot of Lewis's argument can be coherently denied.

Once more, in order for this argument to be convincing, we need to give reasons to think that 'A's truth would be good' really is a context-sensitive sentence. Thankfully, however, those reasons aren't hard to come by. Most obviously, think again about the second statement of the DAB thesis we looked at—namely, the one which made use of an indicative conditional:

$$v(A) = p(A > G). \qquad \text{(Conditional DAB)}$$

As we heard before, many philosophers (and linguists) believe that indicative conditional sentences are context-sensitive. In other words, which proposition is expressed by a sentence like 'If A, then G' depends, in part, on a body of evidence that's salient in the context. However, if that's right, then it simply follows

[21] For more fully developed defenses of this version of contextualism, see, e.g., Stalnaker (1975, 1984), van Rooij (1999), Bacon (2015), Schultheis (2023), Mandelkern (forthcoming), or Dorr and Hawthorne (MS).

from this that the conditional version of the DAB thesis is best viewed as a context-sensitive thesis. After all, which proposition plays the relevant role on the right-hand side of the equation above depends on which proposition 'If A, then G' picks out.

That said, even if you reject the conditional formulation of the DAB thesis, there's still reason to think that 'A's truth would be good' is a context-sensitive sentence—and thus that different propositions can play the A°-role in different contexts. To illustrate, consider an example from Price (1989) (whose dates I've updated):

> My dear Aunt Agatha may die in 2025. Let A be the proposition that she does so. My interest in the truth of A is entirely constrained by the facts that I am Aunt Agatha's sole heir, that she is periodically very wealthy, and that money is my only joy. Thus I think that it would be good if Aunt Agatha dies in 2025 if and only if in that case I inherit a fortune… Both Agatha's prospects and mine depend on the state of the economy. There may or may not be a recession in 2025… [and t]his affects my prospects in this way: I am a lot less likely to inherit if Agatha dies in a recession than if she dies otherwise (since obviously in a recession there is less likely to be a fortune for me to inherit)… As for Aunt Agatha's prospects, she has always said that she would hate to die in a bear market, and so she will try to hang on if the economy is down. (p. 123)

Now, suppose that one morning, while reading the *Financial Times*, I learn that a bear market is predicted for 2025. Then, in this context, it would sound false for me to say the following:

(2) Aunt Agatha's death in 2025 would be good.

On the other hand, if I had learned that a bull market is predicted for 2025, then it seems like an utterance of (2) would've been true, rather than false. So, in much the same way as we saw in Gibbard's "Sly Pete" case, it seems like the proposition expressed by (2) depends on the evidence that's available to me in the context. That is, different propositions are expressed by that sentence depending on which background possibilities are salient. However, if that's right, then we have good reasons to believe that an adequate version of the DAB thesis is context-sensitive, too. And as we saw, on this version of the thesis, Lewis's argument doesn't hold.

3.4 Tenability

We've thus found a way of resisting Lewis's argument against the DAB thesis. On a contextualist reading of that thesis, his argument doesn't hold. What I want to do now, however, is show that we can establish something stronger. Specifically, if we view the DAB thesis in the contextualist way I've suggested, then we can prove a tenability result for that thesis, which shows that no argument analogous to Lewis's can be given.[22]

The tenability result I'll sketch in this section draws on tenability results for Stalnaker's thesis proved by van Fraassen (1976), Bacon (2015), Goldstein and Santorio (2021), and Khoo (2022). In fact, my result is really an application of those results. For this reason, then, I'm going to take the conditional formulation of the DAB thesis as canonical going forward. Recall that this version of the thesis said the following:

$$v(A) = p(A > G). \qquad \text{(Conditional DAB)}$$

[22]Establishing a result like this is important, because there are other arguments against the DAB thesis in the literature, in addition to Lewis's—see, for example, Collins (1988) or Costa et al. (1995). And it's not yet obvious that the contextualist strategy evades these other arguments against the DAB thesis. However, establishing the tenability result for the contextualist DAB thesis shows that even these other arguments don't affect that thesis.

That is, your desire for A should match your credence in the proposition expressed by 'If A, then G' in the relevant context. (Note that this is the contextualist reading of the conditional DAB thesis. That's the interpretation I'll have in mind from this point on. More precisely, whenever I write '$A > G$' in what follows, I'll mean the contextually salient proposition expressed by 'If A, then G'. And the same goes for arbitrary indicative conditionals, $A > C$.)

Now, following the authors I just mentioned, I'm also going to adopt a particular semantics for indicative conditionals to prove my tenability result. Roughly speaking, this semantics is based on that of Stalnaker (1968). On Stalnaker's view, a conditional's truth-value depends on a relation of *similarity* between possible worlds, and the semantics I'll adopt here does the same. I won't say much to *justify* this semantics in what follows—the fact that it can be used to vindicate the DAB thesis is arguably justification enough. However, readers who need more convincing should see the able defenses of this semantics given by Khoo (2022), Schultheis (2023), Mandelkern (forthcoming), or Dorr and Mandelkern (MS). As you'll see if you look at those works, this semantics is motivated on independent, linguistic grounds.

Thus, to spell out the semantics for indicative conditionals, we first need to introduce the notion of a *sequence*. Formally, a sequence is an n-tuple of the worlds in $\mathcal{W}$, without repetitions. (Recall that $\mathcal{W}$ is the set of all your epistemically possible worlds.) Informally, however, it's an *ordering* of the worlds in $\mathcal{W}$, ordered according to how similar they are to the first world. For example, if $\mathcal{W}$ is the set $\mathcal{W} = \{w_1, w_2, w_3\}$, then the sequence $\langle w_1, w_2, w_3\rangle$ says that w_1 is the most similar world to itself, w_2 is the next most similar, and so on.

In the present setting, sequences function as the points of evaluation for indicative conditional sentences. That is, whereas ordinary "factual" sentences are true or false at possible worlds, indicative conditional sentences are true or false *at sequences of worlds*. Specifically, 'If A, then C' is true at a sequence, on this semantics, just in case the first A-world in that sequence is a C-world. The idea here is supposed to be akin to Stalnaker's, that a conditional's truth-value depends on a relation of similarity between possible worlds.

It may be worth pausing here to look at an example. So, consider again the toy case, where $\mathcal{W} = \{w_1, w_2, w_3\}$. Given this set of worlds, there are six possible sequences that we can generate, where these are contained in the set $\mathcal{S}$, written below:

$$\mathcal{S} = \left\{ \begin{array}{l} \langle w_1, w_2, w_3\rangle, \langle w_1, w_3, w_2\rangle, \\ \langle w_2, w_1, w_3\rangle, \langle w_2, w_3, w_1\rangle, \\ \langle w_3, w_1, w_2\rangle, \langle w_3, w_2, w_1\rangle \end{array} \right\}.$$

Now, suppose A is a factual proposition true at w_1 and w_2, and C is a factual proposition true at w_2 and w_3. Then, the indicative conditional $A > C$ is true at three sequences in $\mathcal{S}$—namely, $\langle w_2, w_1, w_3\rangle$, $\langle w_2, w_3, w_1\rangle$, and $\langle w_3, w_2, w_1\rangle$—since these are the only sequences whose first A-world is a C-world.

To see why this is a plausible semantic view about indicative conditionals, let's look at a different example. Recall the sentence (3), which we looked at in §4.2:

(3) If this (fair, six-sided) die landed on an even number, then it landed on 2.

Suppose you're ignorant about how the die landed. And suppose we model your state of ignorance with a set of six worlds, $\mathcal{W} = \{w_1, ..., w_6\}$. Here, each world w_i is a world where the die landed on i (for $i = 1, ..., 6$). Now, at a world like w_3, where the antecedent of (3) is false, is the sentence (3) determinately true, or determinately false? Prima facie, there's no clear way of answering that question. And our semantics for indicative conditionals tells us why. The reason is that the coarse-grained, "descriptive" facts that obtain at w_3—in particular, that the die landed on 3 at that world—don't pin down how the die landed *if* it landed on an even number. For that, we need to posit a relation of similarity between the worlds in $\mathcal{W}$. And that's the role that sequences play in the present construction.

In a moment, we'll see that this "fine-grained" way of thinking about indicative conditionals is really the key to vindicating Stalnaker's thesis, and also my contextualist version of the DAB thesis. In the meantime, however, let me say something about why it's right to think of the semantics just given as a contextualist semantics for indicative conditionals. That much isn't yet obvious. But I think there are two key reasons to think that this is so.

First, recall my initial gloss of contextualism about indicative conditionals from §3.3. There, I said that, according to this view, the proposition expressed by a sentence 'If A, then C' depends on a body of evidence that's available to speakers in the context. One way we can model this body of evidence is as a set of possible worlds—specifically, as the set $\mathcal{W}$ consisting of all the worlds that you, and perhaps others in the conversation, believe could be actual. However, above I said that the sequences we can construct in a given context are drawn from the worlds in $\mathcal{W}$. So, there's a clear sense in which these sequences depend on what evidence is available in the context. In other words, as you learn new information and the context is updated, so too is the set of sequences that we can use to interpret indicative conditional sentences. In turn, exactly what a sentence like 'If A, then C' expresses can change as you learn new information.

The other—and more important—reason for thinking that this semantics is a contextualist semantics, however, is that, for philosophers who endorse contextualism, there isn't just one privileged notion of similarity that's suitable across all contexts. Instead, exactly what we mean by 'most similar A-world' can change depending on the context, with some similarity relations being appropriate in some contexts, and others being appropriate in others. To illustrate this, look again at Price's example from the previous section. In that case, we heard that whether or not I inherit a fortune if Aunt Agatha dies in 2025 depends on the state of the economy. Thus, it seems like, at any world at which there's a recession in 2025, the most similar world at which Aunt Agatha dies should be one at which I don't inherit a fortune. And at any world in which there's not a recession in 2025, it seems like the most similar world at which Aunt Agatha dies should be one which I do inherit a fortune. In other contexts, however, different relations of similarity might be more appropriate—or admissible, as I'll sometimes say. For example, if Aunt Agatha was prone to give up on life during a bear market, rather than to hang on, then the appropriate similarity relations in this context would be the opposite of what we've just seen.

Thus, the sequence-based semantics for indicative conditionals fits very naturally with the contextualist view about those conditionals that I mentioned in previous sections. That said, thinking about indicative conditionals in this new, "fine-grained" way raises an important issue. To see what it is, recall that in §2 I assumed that your credence function, p, is defined only on sets of worlds. However, we're now thinking about indicative conditionals as corresponding to sets of sequences. So, if you're going to assign meaningful credences to these conditionals, then it seems like we need to find a way of extending your credence function, p, so that it's defined over sequences, and not just over worlds.

To make this extension, then, I'm going to follow a suggestion of Goldstein and Santorio (2021) and Khoo (2022), who in turn draw on ideas from van Fraassen (1976). The basic thought is to "lift" your credence function, p, to a new credence function, q, defined over sequences, using a recursive procedure (below, I write '$[w]$' for the set of sequences beginning with w, and '$[w_1, ..., w_k]$' for the set of sequences that share the same k-length initial segment):

(i) $q([w]) = p(w)$,

(ii) $q([w_1, ..., w_k]) = q([w_1, ..., w_{k-1}]) \cdot p(w_k \mid \mathcal{W} - \{w_1, ..., w_{k-1}\})$.

Heuristically, we can think of this as saying that your credence in a sequence, $\langle w_1, ..., w_n \rangle$, is your credence that you'd draw those world from an urn, in that order, and without replacement. When q comes from p in this way, I'll say that q is well-behaved. (Note also that the recursive procedure above requires that p assigns positive credence to each world $w \in \mathcal{W}$. That's why I assumed that p is regular, in the sense of assigning positive credence to each $w \in \mathcal{W}$, way back in §4.2.)

To see how this lifting procedure works in action, let's consider again the toy example, where $\mathcal{W} = \{w_1, w_2, w_3\}$, and the six admissible sequences are those I listed above. Let's also assume that A is the proposition $\{w_1, w_2\}$, and C is the proposition $\{w_2, w_3\}$. Suppose also that $p(w_1) = 1/2$, $p(w_2) = 1/3$, and $p(w_3) = 1/6$. Then it follows that $p(A) = 5/6$, $p(C) = 1/2$, and—importantly—$p(C \mid A) = 2/5$. Note that the value of this conditional credence will be very important in a moment.

Now, it's easy to show that a credence function q, which extends p according to (i) and (ii), preserves all of these values just given. In particular, if q is defined according to (i) and (ii), then $q(A) = 5/6$, $q(C) = 1/2$, and $q(C \mid A) = 2/5$. Strikingly, however, q also assigns the value $2/5$ to the indicative conditional $A > C$. To see why, just consider our semantics. According to that semantics, there are three sequences for which the first A-world is also a C-world—as we heard—where these are: $\langle w_2, w_1, w_3\rangle$, $\langle w_2, w_3, w_1\rangle$, and $\langle w_3, w_2, w_1\rangle$. But if we appeal to the lifting procedure (i) and (ii) above, it follows that:

$$\begin{aligned} q(\langle w_2, w_1, w_3\rangle) &= q([w_2, w_1]) \cdot p(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= q([w_2]) \cdot p(w_1 \mid \mathcal{W} - \{w_2\}) \cdot p(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= p(w_2) \cdot p(w_1 \mid \mathcal{W} - \{w_1\}) \cdot p(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= 1/3 \cdot 3/4 \\ &= 1/4. \end{aligned}$$

Similar calculations show that $q(\langle w_2, w_3, w_1\rangle) = 1/12$, and $q(\langle w_3, w_2, w_1\rangle) = 1/15$. And if we take the sum of your credences in all of these sequences, we get that $q(A > C) = q(C \mid A) = 2/5$. What we have here, then, is a proof that Stalnaker's thesis can be satisfied non-trivially in the present setting. Contrary to Lewis's triviality results, we can construct a case in which that thesis holds non-trivially after all.

Better still, suppose that C just is the proposition G, consisting of all and only "the objectively desirable [worlds]—for short, the good ones" (Lewis, 1996, p. 307). Then, by Jeffrey's definition of rational subjective value, we have that $v(A) = q(G|A)$. But also, by the result about Stalnaker's thesis that we just saw, we have that $v(A) = q(A > G)$. So the foregoing example is a case in which the DAB thesis holds non-trivially, too. (And note that Invariance is also satisfied in this example, as the reader can check for themselves. This follows since, in the case in question, $q(A > G) = q(A > G|A)$.)

This, it turns out, is an instance of a much more general result, proved in the Appendix. (Note also that, in the more general case, the tenability of DAB doesn't simply reduce to establishing the tenability of Stalnaker's thesis.) That result says—so long as q is a well-behaved credence function, and as long as the all sequences of worlds are admissible similarity orderings in the context—that the DAB thesis holds even in cases in which we allow for a range of propositions about goodness. More precisely, we have the following general result:

Theorem 1 (Generalized DAB Thesis)**.** Let q be a credence function that extends p, defined according to the recursive procedure (i) and (ii) above. Let v be a subjective value function, defined in line with Jeffrey's theory. Let $\{G_x\}$ be a partition of propositions, each member of which says *the world is good to degree* x. Then, if the context is transparent:

$$v(A) = \sum_x q(A > G_x) \cdot x. \qquad \text{(Generalized DAB)}$$

In other words, a generalized version of the DAB thesis holds.[23]

The simpler version of the conditional DAB thesis that we've mostly been focusing on in this paper is a special case of the Generalized DAB thesis above. Specifically, it's the special case in which $\{G_x\}$ contains

[23]See Lewis (1988a, p. 330) for a more general statement of the DAB thesis, akin to this one.

only two propositions, namely G and $\neg G$. Thus, the upshot of this result is that, provided we think of the DAB thesis in the context-sensitive way I've set out in this section, that thesis is compatible with Jeffrey's decision theory after all. Not only does Lewis's argument against that thesis not succeed, but no other, analogous argument can even been given.

3.5 Comparison

The tenability result for the DAB thesis I've just laid out should come as good news to those of us who think that this thesis is plausible. That result is, however, superficially similar to a couple of other responses to Lewis's argument that have appeared in the philosophical literature. In this section, then, I want take a bit of time to compare my result to those others. As we'll see, both of these responses have a lot going for them. But I nevertheless think that there are strong reasons to prefer the contextualist response that I sketched in §§3.3–3.4.

3.5.1 Bradley and List's Response

I'll start with a response to Lewis's argument given by Bradley and List (2009). In their paper, Bradley and List say that we can give a version of the DAB thesis that coheres with Jeffrey's decision theory, provided we're willing to distinguish between "purely factual" and "purely evaluative" propositions. To see how this response works, first suppose that every world, w, can be decomposed into a pair, $\langle f, g\rangle$, where f is a maximally specific "factual state", and g is a maximally specific "evaluative state". As Bradley and List say, "we can think of f as capturing "the totality of physical facts holding in [a] world, and g the totality of normative facts (e.g. ought facts or goodness facts)" (p. 33–34; with trivial changes of notation). In turn, logical space can be thought of as the set of all such factual state/evaluative state pairs.

Now, once we have this picture of logical space in mind, we can define 'purely factual proposition' as follows. First, let $\mathcal{F}$ be the set of all maximially specific factual states, and let $\mathcal{G}$ be the set of all maximally specific evaluative states. Then, a *purely factual proposition*, A, is a proposition of the form $A_{\mathcal{F}} \times \mathcal{G}$, where $A_{\mathcal{S}} \subseteq \mathcal{F}$. We can then define 'purely evaluative proposition' similarly. A *purely evaluative proposition*, A°, is a proposition of the form $A^{\circ}_{\mathcal{G}} \times \mathcal{F}$, where $A^{\circ}_{\mathcal{G}} \subseteq \mathcal{G}$.

Bradley and List then show that it suffices to satisfy a version of the DAB thesis—as well as the other assumptions that Lewis makes—that your credence in any world, w, be equal to the *product* of your credences in the factual state f and the evaluative state g, of which w is composed. That is, for each world w, we should have $p(w) = p(f) \cdot p(g)$. To see why this leads to satisfaction of the DAB thesis, first let A be a purely factual proposition, and let A° be a purely evaluative proposition. Then, since each world w in the conjunction $A \wedge A^{\circ}$ is such that $p(w) = p(f) \cdot p(g)$, it follows that $p(A \wedge A^{\circ}) = p(A) \cdot p(A^{\circ})$. In other words, A and A° are probabilistically independent. And from this it follows that conditionalizing on the truth of A does nothing to change your credence in A°: $p(A^{\circ}) = p(A^{\circ} \mid A)$.

Now, let's define a subjective value function v over non-evaluative propositions only, and set $v(A) = p(A^{\circ})$. Then, the DAB thesis and Invariance are both satisfied. In the first case, DAB is satisfied by construction. And in the second case, we have that $v_A(A) = v(A)$, since $p_A(A^{\circ}) = p(A^{\circ} \mid A) = p(A^{\circ})$, as we just saw. Thus, it looks like Lewis's argument has been overcome.

Now, like I said, this result is in some ways superficially similar to mine. In particular, it appeals to a similar technique, taking worlds to be entitied capable of fine-graining. Unfortunately, however, Bradley and List's way of circumventing Lewis's argument, based on this idea, is not fully convincing. The main issue is that, in taking logical space to be composed of *all* factual state/evaluative state pairs, Bradley and List require us to think that any factual state can be "freely recombined" with any evaluative state. In turn, this requires us to deny a near-universally accepted meta-ethical thesis—namely, that the evaluative facts at a world *supervene* on the non-evaluative facts. (Actually, Bradley and List's argument requires something

even stronger—namely that the evaluative facts aren't even *partly determined* by the non-evaluative facts.) As Rosen (2017) puts it, however, "The idea that there cannot be two actions that are alike in every non-normative respect, one of which is right and the other wrong is as close to common ground as we get in metaethics" (pp. 153-54). Thus, if Bradley and List's way of vindicating the DAB thesis requires us to deny supervenience—and even partial determination—then this seems like a good reason to reject their approach.[24]

3.5.2 Hájek and Pettit's Response

A very different response to Lewis is given Hájek and Pettit (2004). According to them, we can resist the argument against the DAB thesis by re-jigging the order of the quantifiers in the statement of that thesis.

To see how this works, let me first state the DAB thesis more carefully than I did in §4.2. What that thesis says is that, for all propositions A, there exists a proposition A° such that, for all rational credence function/subjective value function pairs, the following holds:

$$v(A) = p(A^\circ) \tag{DAB}$$

Thus, there are three quantifiers here to deal with: a universal quantifier, followed by an existential quantifier, followed by another universal quantifier.

Hájek and Pettit argue that we can block Lewis's argument against the DAB thesis, provided we reverse the order of the final two quantifiers. That is, rather than thinking of A° as being independent of p and v, we can make A° dependent on these functions. Thus, on this revised statement of the DAB thesis, we have: for every proposition A, and for all rational credence function/subjective value function pairs, there exists a proposition A° such that $v(A) = p(A^\circ)$. And given *this* statement of the DAB thesis, Lewis's argument doesn't go through. As Hájek and Pettit say:

> [This statement of the thesis] evades all the negative results that we have discussed or mentioned, since they assumed that A° remained fixed throughout redistributions of credence... [I]f instead we allow the identity of A° to change as the distribution of probability changes, we have no guarantee that the required cleavages will take place. (2004, p. 85)

Once again, this response to Lewis has some similarities with the response I gave in §§3.3–3.4. In particular, it allows the identity of A° to change, just as I did. However, I think there's a problem with the response, as Hájek and Pettit state it. Specifically, it looks like Hájek and Pettit's version of the DAB thesis undercuts some of that thesis's original motivation. After all, recall that in §4.2 I said DAB seems especially plausible if we're *realists* about value. Oddie (2001), for example, made a case for this claim. And Lewis thinks that the same thing is true. In his 1996, for example, he says—in somewhat exaggerated terms—that DAB promises a

> rich reward: objective ethics. If there are some things we desire by [rational] necessity, we surely would want to say that these things were objectively desirable. Or if there were some propositions, belief or disbelief in which was [rationally] connected with desire, some of them presumably would be true; then we surely would want to say that the true ones were the objective truth about ethical reality. (p. 307)

As Hájek and Pettit state the DAB thesis, however, this realist motivation seems to be lost. Specifically, by swapping the order of the quantifiers in the statement of that thesis, they make A° "indexical", in the sense that its truth or falsity *depends* on your credences and values.

[24] Let me note here that Bradley and List anticipate this objection, and argue that, if evaluative facts really do supervene on non-evaluative facts, then Lewis's Invariance assumption isn't warranted. However, in the next section, I'll say more about why I think this response isn't right.

This, in fact, is how Hájek and Pettit sell their argument. In their discussion, they say that one appropriate meta-ethical theory which their version of the DAB thesis fits with is "that G. E. Moore… called *subjectivism*. It holds that when someone says that a prospect is good, then that utterance expresses the belief that the speaker has an attitude of approval towards the prospect" (p. 86). Later on, they tie their version of the DAB thesis to more sophisticated "expressivist" views about the nature of value. But the point remains the same: while this version of the DAB thesis avoids Lewis's objection, it's not in the spirit of the original DAB thesis. On the contrary, the reason that I and many others were attracted to that thesis in the first place is that it provides a putative rational link between your subjective values, on the one hand, and your beliefs about the objective values, on the other. So, if a revised version of the thesis doesn't provide us with a similar rational link, or requires us to give up on the notion of objective value altogether, then my feeling is that this is a reason to reject it.[25]

Note also that my contextualist version of the DAB thesis doesn't fall prey to this objection. While it's true to say that, like Hájek and Pettit's version of the thesis, the contextualist DAB thesis allows the identity of A° to change, the *way* in which it changes doesn't depend your credences or desires. Instead, the proposition expressed by a sentence like 'If A, then G' depends on the context, on my view. And this needn't have anything to do with your credences or values. Thus, unlike the "indexical" DAB thesis, the contextualist DAB thesis is compatible with that thesis's realist motivations.

3.6 When the DAB Thesis Fails

I've now made my case for the contextualist DAB thesis. As I said, I think there are strong reasons to prefer that thesis to other versions that philosophers have offered. At the same time, however, I don't think it's *completely* right, as we'll now see. And I want to close now by discussing some of that thesis's limitations.

To start off, then, recall what my tenability result from §3.4 established. That result showed that the DAB thesis holds in every context, in which every sequence of possible worlds counts as an admissible similarity ordering. As it turns out, however, something even more general is true. This is that the DAB thesis is satisfied in any context in which a certain "Independence condition" holds.

To see what I mean, let me first state a fact about the semantics for indicative conditionals that I worked with in §3.4. This is that it validates a principle known as *Probabilistic Centering*:[26]

$$q(A \wedge (A > C)) = q(A \wedge C). \qquad \text{(Probabilistic Centering)}$$

In words, Probabilistic Centering says that your credence in a conditional, together with its antecedent, is equal to your credence in the conjunction of antecedent and consequent. And intuitively, that seems right. (Just think, for instance, about how confident you are in the following sentences: 'The die landed even and if it landed even, it landed on 2' and 'The die landed even and it landed on 2'.)

Now, interestingly, Probabilistic Centering *implies* Stalnaker's Thesis, whenever your credences in a conditional $A > C$ and its antecedent are independent of one another. That is, whenever the following condition holds:

$$p(A > C) = p(A > C \mid A), \qquad \text{(Antecedent Independence)}$$

[25]Bacon (2015) makes a somewhat similar objection to van Fraassen's (1976) indexical version of Stalnaker's thesis. He says that: "There does not appear to be an independently motivated reason to think that two conditional utterances, made when the same epistemic possibilities are open, could express different propositions due to a small difference in how probable these possibilities are [something that's allowed by van Fraassen's version of Stalnaker's thesis]. This seems like a fairly radical form of context sensitivity" (p. 140). I agree. And in my view, similar objections apply to the Hájek-Pettit version of the DAB thesis.

[26]Probabilistic Centering follows from the fact our semantics validates the logical principles known as *Weak Centering* and *Strong Centering*. See Rothschild (2011) for further discussion.

then our semantics implies that Stalnaker's thesis holds also.[27] My contextualist DAB thesis follows from Stalnaker's thesis, as we saw (at least in the very simplest of cases). So, what this means is that, whenever Antecedent Independence is satisfied, the DAB thesis holds as well.

But is *every* context one in which Antecedent Independence is satisfied? Some philosophers think this is the case.[28] But I think there are strong reasons to doubt. Indeed, Price's case, which we looked at in §3.3, seems like a counterexample to this claim. To see why, recall that in that case, I stood to inherit a fortune from Aunt Agatha's death if and only if she didn't die in a recession. But at the same time, Aunt Agatha's death in 2025 would provide me with excellent evidence that there's *not* a recession looming, since "she has always said that she would hate to die in a bear market, and so she will try to hang on if the economy is down" (Price, 1989, p. 122). Thus, it seems like my conditional credence that things will be good, given that Aunt Agatha dies in 2025, should be high in this scenario— with the reason being that learning of her death would give me strong evidence about the state of the economy. However, this doesn't match my intuitive credence in the conditional below:

(3) If Aunt Agatha dies in 2025, then things will be good.

In this case, in contrast, my credence in (3) is middling, since I'm unsure if there will be a recession in 2025. Thus, it looks like we have a violation of Stalnaker's thesis—and *a fortiori*, we have a violation of the DAB thesis as well.

We can explain this failure of the DAB thesis by noting that Price's case is one in which there's a failure of Antecedent Independence. To see this, let A be the proposition that Aunt Agatha dies in 2025. Then, as I said just said, my credence $p(A > G)$ is middling. However, my credence $p(A > G \mid A)$ is high, since *learning* that Aunt Agatha has died would provide me with strong evidence that things are good. (In fact, probabilistic centering implies here that my credence $p(A > G \mid A)$ just *is* my conditional credence $p(G \mid A)$—it's an easy exercise to show that).

Now, if this is right, and some contexts are such that Antecedent Independence fails, then there are a couple of things we need to say about it. The first is that cases like this seem to give us a prima facie reason to that Lewis's Invariance assumption isn't warranted. After all, recall that Invariance—which played a key role in the argument in §4.2—said that $v_A(A) = v(A)$, or (informally) that your subjective value for A doesn't change when you learn that A. However, if your credence in $A > G$ can change when you come to learn that A, then there's reason to think that Invariance is invalid. As Bradley and List (2009) state this objection, for example:

> If the independence requirement is violated... it is simply not clear whether we still have any reason to insist on the invariance requirement. If there is a correlation between A and $A > G$, then it is no surprise that our evaluation of A may change after learning that A. (p. 36, with trivial changes of notation)

But this, it turns out, isn't right. Contrary to what Bradley and List say here, the Invariance assumption is *implied* by Jeffrey's theory—and so, in a sense, it's not a really an extra assumption Lewis had to make in the first place. (I mentioned this in passing way back in §4.2.) To see this, recall that in Jeffrey's theory,

[27] *Proof.*

$$\begin{aligned} p(A > C) &= p(A > C \mid A) && \text{(Independence of } A \text{ and } A > C) \\ &= p(A \wedge (A > C))/p(A) && \text{(Def. of Conditional Probability)} \\ &= p(A \wedge C)/p(A) && \text{(Probabilistic Centering)} \\ &= p(C \mid A) && \text{(Def. of Conditional Probability)} \end{aligned}$$

□

[28] See, e.g., McGee (1989), Bradley (2012), or Goldstein and Santorio (2021).

$v(A) = p(G \mid A)$. But the definition of Conditionalization implies that $p_A(G \mid A) = p(G \mid A)$.[29] In turn, this implies that $v_A(A) = v(A)$. And thus failures of Antecedent Independence don't imply failures of Invariance, contrary to what Bradley and List claim.[30]

At the same time, however, failures of Independence do imply failures of the DAB thesis, as I mentioned just a moment ago. Indeed, this is precisely the point that Price was trying to make with his example. In explaining that case, he says that:

> if we can see that the value we ascribe to A would be liable to change, *were we to discover that* A, then the appropriate value to use in deliberation is the value A *would have for us* in those circumstances [i.e., $v(A) = p(G \mid A)$]. The guiding principle is that whenever it makes a difference, we should assess a possible outcome under the hypothesis that it is the actual outcome. (p. 122; emphasis in the original)

Thus, when Antecedent Independence is violated, the DAB thesis doesn't hold.

I think Price is right about this. What's more, I think it has important implications for how we should view the DAB thesis. I started this paper by saying that DAB is a simple and compelling thesis about the relationship between your subjective values, on the one hand, and your beliefs about the objective values, on the other. What the foregoing shows, however, is that the thesis is probably *too* simple. When there are correlations between your credence in a proposition about A's goodness, and your credence in the proposition A itself, that thesis doesn't capture the correct way in which your beliefs and desires relate. Rather, in those sorts of cases, a more complicated relationship obtains. And that relationship is captured by Jeffrey's theory.

But at the same time, that doesn't mean the DAB thesis is without merit. On the contrary, there are still lots of contexts in which it holds. Indeed, that's what my tenability result establishes. It shows that, if we view the DAB thesis in the contextualist way I've suggested, then—contra Lewis—that thesis often holds non-trivially after all.

Appendix

In this Appendix, I prove Theorem 2, stated in the main text. The proof is essentially an application of the proof of Theorem 1 in Goldstein and Santorio (2021). See also Khoo and Santorio (2018, Chapter 5) and Khoo (2022, pp. 161-62).

Proof. Since $\{G_x\}$ is a partition, Jeffrey's definition of subjective value implies that, for any proposition A:

$$v(A) = \sum_x q(G_x \mid A) \cdot v(A \wedge G_x)$$

Thus, there are two things that we need to show here:

(A) for each G_x, $v(A \wedge G_x) = x$,

(B) for each G_x, $q(G_x \mid A) = q(A > G_x)$.

[29] This is known as the *rigidity property* of Conditionalization. We can prove it as follows:

$$p_A(C \mid A) = \frac{p(A \wedge C \mid A)}{p(A \mid A)} = \frac{p(A \wedge C \wedge A)/p(A)}{p(A \wedge A)/p(A)} = \frac{p(A \wedge C \wedge A)}{p(A \wedge A)} = \frac{p(C \wedge A)}{p(A)} = p(C \mid A).$$

[30] Incidentally, this fact is noted by Bradley himself in a later paper. See Bradley and Stefánsson (2016).

Let's start with (A). First, fix an arbitrary $G_x \in \{G_x\}$. Then, since $A \wedge G_x$ is a subset of G_x, each world w in $A \wedge G_x$ is such that $v(w) = x$. It follows that:

$$\begin{aligned} v(A \wedge G_x) &= \sum_w q(w \mid A \wedge G_x) \cdot v(w) \\ &= \sum_w q(w \mid A \wedge G_x) \cdot x \\ &= x. \end{aligned}$$

Then, since our choice of G_x was arbitrary, this establishes (A).

Now turn to (B). Again, fix a particular G_x, and let $\mathcal{S}$ be the set of all sequences we can generate from $\mathcal{W}$. We partition $\mathcal{S}$ into equivalence classes, $\mathcal{S}_A^1, ..., \mathcal{S}_A^n$, where each $\mathcal{S}_A^i$ is the set of sequences whose first A-world occurs in position i. By the law of total probability:

$$q(A > G_x) = q(A > G_x \mid \mathcal{S}_A^1) \cdot p(\mathcal{S}_A^1) + ... + q(A > G_x \mid \mathcal{S}_A^n) \cdot q(\mathcal{S}_A^n).$$

So, what we'll now show is that, for each i, $q(A > G_x \mid \mathcal{S}_A^i) = q(G_x \mid A)$. Then, since $\sum_i q(\mathcal{S}_A^i) = 1$, this will establish that $q(A > G_x) = q(G_x \mid A)$.

Thus, consider an arbitrary set $\mathcal{S}_A^i$. By the ratio formula:

$$q(A > G_x \mid \mathcal{S}_A^i) = \frac{q((A > G_x) \wedge \mathcal{S}_A^i)}{q(\mathcal{S}_A^i)}.$$

So, what we need to do is find the credences in the numerator and denominator on the right-hand side.

Starting with the former: first note that the conjunction $(A > G_x) \wedge S_A^i$ is the set of sequences whose first A-world occurs in position i, and where G_x is true at that world. Thus, to find the credence that pr assigns to this set, we appeal to the lifting procedure from §3.3. Doing so, we get:

$$\begin{aligned} q((A > G_x) \wedge \mathcal{S}_A^i) &= \sum_{w_1,...,w_{i-1} \in \neg A} \sum_{w_i \in A \wedge G_x} q([w_1, ..., w_{i-1}, w_i]) \\ &= \sum_{w_1,...,w_{i-1} \in \neg A} \sum_{w_i \in A \wedge G_x} q([w_1, ..., w_{i-1}]) \cdot \frac{p(w_i)}{p(W - \{w_1, ..., w_{i-1}\})}. \end{aligned}$$

By similar reasoning:

$$\begin{aligned} q(\mathcal{S}_A^i) &= \sum_{w_1,...,w_{i-1} \in \neg A} \sum_{w_i \in A} q([w_1, ..., w_{i-1}, w_i]) \\ &= \sum_{w_1,...,w_{i-1} \in \neg A} \sum_{w_i \in A} \cdot \frac{p(w_i)}{p(W - \{w_1, ..., w_{i-1}\})} \end{aligned}$$

So all we need to do now is solve for $\frac{q((A > G_x) \wedge \mathcal{S}_A^i)}{q(\mathcal{S}_A^i)}$. Here, most of the terms cancel, leaving us with:

$$\begin{aligned} \frac{q((A > G_x) \wedge \mathcal{S}_A^i)}{q(\mathcal{S}_A^i)} &= \sum_{w \in A \wedge G_x} \sum_{w' \in A} \frac{p(w)}{p(w')} \\ &= \frac{p(A \wedge G_x)}{cr(A)} \\ &= p(G_x \mid A). \end{aligned}$$

Then, since q and p agree about the credences assigned to factual propositions, it follows that $q(G_x \mid A) = p(G_x \mid A)$. Thus we've established that $q(A > G_x \mid \mathcal{S}^i_A) = q(G_x \mid A)$.

Now, since our choice of $\mathcal{S}^i_A$ was arbitrary, the same conclusion follows for any $\mathcal{S}^i_A$. So, by the law of total probability:

$$\begin{aligned} q(A > G_x) &= q(A > G_x \mid \mathcal{S}^1_A) \cdot q(\mathcal{S}^1_A) + ... + q(A > G_x \mid \mathcal{S}^n_A) \cdot q(\mathcal{S}^n_A) \\ &= q(G_x \mid A) \cdot pr(\mathcal{S}^1_A) + ... + q(G_x \mid A) \cdot pr(\mathcal{S}^n_A) \\ &= q(G_x \mid A). \end{aligned}$$

Finally, since our choice of G_x was arbitrary, the foregoing holds for any G_x. Thus, combining (A) and (B) we get that:

$$v(A) = \sum_x q(A > G_x) \cdot x,$$

which is what we were trying to show. □

Chapter 4

Learning 'If'

4.1 Introduction

Private Judy Benjamin—that one-time prosperous Brooklynite, who's been unwittingly recruited to the army—has just been dropped into unfamiliar territory with her platoon.[1] The territory is divided into two halves: a Red Territory and a Blue Territory. Each of these halves is divided into two further halves: a Headquarters Company Territory and a Second Company Territory. Thus, Judy is in an area with four quadrants, equally-sized:

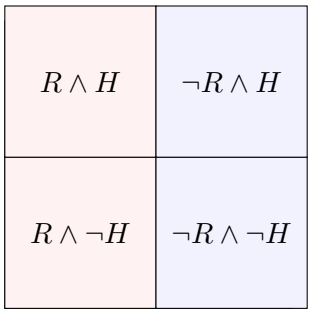

Figure 1. Judy's Predicament

Judy has no idea where she is in this territory. So she spreads her (prior) credences evenly over the various possibilities. If p is her (prior) credence function (subjective probability function), R is the proposition that she's in Red Territory, and H is the proposition that she's in Headquarters Company Territory, then Judy's credences are:

$$p(R \wedge H) = p(R \wedge \neg H) = p(\neg R \wedge H) = p(\neg R \wedge \neg H) = 1/4.$$

After some time, Judy's Captain appears on the radio. She describes her location to him, and he replies. "I've got no idea where you are", the Captain says. "But:

(1) The probability is $3/4$ that if you're in Red Territory, then you're in Headquarters Company Territory."[2]

[1] This example is from van Fraassen (1981), inspired by the 1980 movie *Private Benjamin*, starring Goldie Hawn.

[2] Here, I'll be neutral about what kind of probability is involved in the Captain's statement. It could, for example, be objective chance, the probability on the Captain's evidence, or something else.

At that moment, the radio crackles and dies. How should Judy's credences change when she hears the Captain utter the indicative conditional (1)?[3]

A few things seem clear. First, it seems like Judy's conditional credence that she's in Headquarters Company Territory, given that she's in Red Territory, should now be $3/4$. Formally, where q is Judy's posterior credence function, after hearing the Captain's testimony:

$$q(H \mid R) = 3/4. \tag{J1}$$

At the same time, it seems like Judy's unconditional credence that she's in Red Territory should remain unchanged. In other words, we should have:

$$p(R) = q(R) = 1/2. \tag{J2}$$

After all, it doesn't seem like the Captain's utterance gives Judy any unconditional information about her location.

Finally, it seems like, for any proposition, and for any $X \in \{R \wedge H, R \wedge \neg H, \neg R\}$, Judy's credence in that proposition, conditional on X, should remain the same as it was before. More precisely, we should have:

$$p(- \mid X) = q(- \mid X), \tag{J3}$$

with the idea being that the Captain's utterance of (1) doesn't seem like it should prompt any change in Judy's conditional credences, except the change in $p(H \mid R)$ (and thus also in $p(\neg H \mid R)$, since Judy satisfies the probability axioms).

All of these conditions are intuitive. However, as van Fraassen (1981) famously claimed, if Judy's posterior credences satisfy these desiderata—and if she doesn't learn anything stronger than what I've just said—then she *can't* be updating those credences in accordance with the standard Bayesian update rules (where by 'standard Bayesian update rules' I mean conditionalization and Jeffrey conditionalization—see below).[4] Faced with this result, some philosophers have concluded that the standard Bayesian update rules can't handle cases like van Fraassen's. The upshot is that Bayesianism isn't nearly as general a theory of rational credence change as many authors claim it is.

The so-called *Judy Benjamin problem* illustrates a more wide-spread problem for Bayesianism. In general, Bayesian update rules seem to give counter-intuitive verdicts—or else fall silent altogether—in *many* cases involving indicative conditionals—i.e., propositions like that expressed by the Captain's sentence (1).[5,6] This leads some authors to lament. As Skyrms (1980a) says, for example, it seems like Bayesians "have no clear conception of what it might be to [update] on a conditional" (p. 169). Similarly, Douven (2012) says that "updating on conditionals [seems to be] very different from standard Bayesian updating" (p. 240). And Eva et al. (2019) say that "Although [indicative conditionals] appear to play a central role in

[3]Indicative conditionals are 'If… then…' statements in the indicative mood. They're often contrasted with *subjunctive* or *counterfactual* conditionals, like: 'If you *were* in Red Territory, then you *would be* in Headquarters Company Territory'. For more on the distinction between these kinds of conditionals, see, e.g., Edgington (1995) or Bennett (2003).

[4]Strictly speaking, van Fraassen argued for something more general. Specifically, he claimed that if Judy's posterior credences satisfy the desiderata (J1)–(J3), then she can't be updating in accordance with any rule that involves a plausible form of "divergence minimization". However, both conditionalization and Jeffrey conditionalization can be viewed as updating methods that involve divergence minimization. Both of these update rules, for example, minimize the divergence between the prior and the posterior, when divergence is measured using something called "relative entropy". (See, e.g., van Fraassen (1981), or Diaconis and Zabell (1982), for further details.) Given this fact, then, I'll focus on van Fraassen's claim as it pertains to the Bayesian updating rules specifically in this paper.

[5]See, e.g., Goldstein and Santorio (2021), Ciardelli and Ommundsen (2022), Fusco (2022), and McNamara and Zhang (MS) for discussions of other cases in this vein.

[6]Note that it's controversial whether indicative conditional sentences express propositions at all. I'll say more about this at the end of this section.

logical and uncertain reasoning… the relationship between [them] and the norms of Bayesian epistemology remains largely opaque" (p. 461).

In this paper, my aim is to make that relationship more transparent.

In what follows, I'm going to show—contrary to what many authors believe—that the standard Bayesian update rules deliver the intuitively correct results in cases like *Judy Benjamin*. To do this, I'll draw a connection between that case, on the one hand, and the notorious thesis known as *Stalnaker's thesis*, on the other (Stalnaker, 1970). Stalnaker's thesis relates your credences in indicative conditionals to your conditional credences. And for a long time it was thought to be untenable, owing to the famous *triviality results* of Lewis (1976) and others. However, recent work has shown that, given a particular semantics for indicative conditionals—a *sequence semantics* which, ironically, was first developed by van Fraassen himself—this thesis is tenable after all. Here I adopt the same semantics in order to rebut van Fraassen's observations in *Judy Benjamin*. Specifically, I show that, given this semantics, the standard Bayesian update rules satisfy all of the intuitive desiderata in that case, contrary to what van Fraassen claimed. I then show that alternatives to the Bayesian update rules, intended to handle cases like *Judy Benjamin*, actually turn out to be equivalent to those rules, according to this semantics—at least in many contexts. Thus, what we end up with is a nice, unified account of rational learning: one which fits well with recent work on the semantics of conditionals, and on Stalnaker's thesis more specifically.

In §4.2, I'll lay the groundwork for my view by discussing two responses to *Judy Benjamin* that have appeared in the philosophical literature. In §4.3 I'll draw out the connection between that case, on the one hand, and Stalnaker's thesis, on the other. I'll use that connection to motivate the sequence semantics for indicative conditionals. And I'll say what updating looks like, given this semantic view. In §4.4, I'll return to *Judy Benjamin* and show that, on the theory of updating sketched in §4.3, all of van Fraassen's desiderata are satisfied. Finally, in §4.5 I'll discuss some of the more general results that can be attained from my theory of updating. §4.6 concludes the paper. And technical details can be found in the appendices.

Before we get started, let me make a few background assumptions clear.

First, I'll assume throughout that your credences at any time satisfy the probability axioms—and so do Judy's—and thus they can always be represented by a probability function, which I'll call your *credence function*. I'll write 'p' for your prior credence function, before a learning experience, and 'q' for your posterior credence function, after a learning experience. I'll write '$A \to C$' for the proposition expressed by the indicative conditional sentence 'If A, then C'. And *pace* authors like Adams (1975), Gibbard (1981), Edgington (1995), and others, I'll assume that indicative conditionals really *do* express propositions—they're not just expressions of, e.g., your conditional credences.[7] (That said, what I *mean* by 'proposition' is arguably a bit non-standard, as we'll see. I'll start by assuming that propositions are sets of worlds. But later on, I'll revise this.) I'll sometimes write 'A, C' in place of '$A \wedge C$', to ease notation. And for the purposes of the paper, I'll focus exclusively on *simple* conditionals—i.e., conditionals that have non-conditional propositions as antecedents and consequents. Complex conditionals present additional challenges, which I'll leave for future research. We'll have enough on our plate just with simple conditionals here.

Lastly, just so we have them in front of us, let me set out the familiar Bayesian rules of rational credence change. The first is:

> **Conditionalization**. After learning a proposition A (and nothing stronger) with certainty, your new credence in any proposition should be equal to your old conditional credence in that proposition, given A. Formally:
>
> $$q(-) = p(- \mid A) := \frac{p(- \wedge A)}{p(A)}, \tag{Cond}$$

[7] See also Bennett (2003), Moss (2015, 2018), and Ciardelli and Ommundsen (2022) for defenses of this "expressivist" view about indicative conditionals.

provided that $p(A) > 0$. (If $p(A) = 0$, then $p(- \mid A)$ is undefined. However, for present purposes, I'll assume that all conditional probabilities under discussion are defined.)

The second rule is:

> **Jeffrey Conditionalization** (R. C. Jeffrey, 1965, 1983). Let $\mathcal{A} = \{A_1, ..., A_n\}$ be a partition of propositions—i.e., a set of propositions that are mutually exclusive and jointly exhaustive. Suppose your credences in the elements of this partition shift so that, for some A_i, $p(A_i) \neq q(A_i)$. Then your new credence in any proposition should be:
>
> $$q(-) = \sum_i p(- \mid A_i) \cdot q(A_i). \tag{J-Cond}$$

It's easy to see that conditionalization is just the special case of Jeffrey conditionalization in which, for some partition element A_i, $q(A_i) = 1$.

4.2 The Lay of the Land

There are two common responses to the *Judy Benjamin* problem in the literature. The first says that we should reject some of van Fraassen's intuitive desiderata, (J1)–(J3). The second says that we should supplement standard Bayesianism, e.g., by introducing additional rules of rational credence change. In this section, I'm going to look at responses of both kinds. I don't think either kind of response fully succeeds. But as we'll see, getting clear on why that is motivates the response that I'll give in later sections.

Thus, to start off, note that authors who make the first kind of response generally focus on van Fraassen's second desideratum, (J2). For example, J. Joyce (2004) singles out this desideratum for special attention, saying that "for all its a priori appeal, (J2) is incorrect. The intuitions in its favor rest on the [erroneous] claim that hearing the Captain say (1) cannot convey any information to Judy about her Red/Blue location" (p. 455; with trivial changes to fit my version of the example). Similarly, Bovens (2009) says that the reasoning supporting (J2) "turns out to be fallacious" (p. 26). And even van Fraassen seems to deny (J2) in later work. In a subsequent, co-authored paper, for instance, he and his collaborators say that Judy should become *less* confident that she's in Red Territory after hearing the Captain's utterance (van Fraassen et al., 1986, p. 455).

For purposes of illustration, let's focus on the argument given by van Fraassen and his collaborators. To see how it works, first consider an alternative version of *Judy Benjamin*, in which everything is the same as it was in the original case, except the Captain now says the following:

(2) If you're in Red Territory, then you're in Headquarters Company Territory.

Thus, rather than telling Judy that there's merely a high probability that she's in Headquarters Company Territory, if she's in Red territory, the Captain now implies that this is a certainty.

If we adapt desideratum (J1) accordingly, then it seems like Judy's conditional credence in H given R should now be 1:

$$q(H \mid R) = 1. \tag{J1*}$$

But if that's right, then $q(R \wedge \neg H) = 0$, by the ratio definition of conditional probability. However, $q(R \wedge \neg H) = 0$ implies that $q(\neg R \vee H) = 1$. And $\neg R \vee H$ is truth-functionally equivalent to the *material* conditional $R \supset H$. Thus, the upshot seems to be that, if hearing the Captain say (2) prompts the change $q(H \mid R) = 1$ in Judy's conditional credences—and if this learning experience doesn't provide her with any additional information—then Judy should change her credences by conditionalizing on the material conditional $R \supset H$.

This is surprising. After all, it's widely agreed that indicative conditionals are not equivalent to material conditionals.[8] Still, the foregoing derivation seems to suggest that the *learnability* conditions for these conditionals are the same. In other words, *learning* an indicative conditional seems to be nothing more than learning a corresponding material conditional.

I'll return to this point in just a moment. For now, however, let's see why it gives us reason to think that desideratum (J2) is flawed. The basic idea is that, in the extreme case in which Judy hears the Captain say (2), (1), conditionalizing on the material conditional $R \supset H$ gives:

$$q(R) = p(R \mid R \supset H) = \frac{p(R, R \supset H)}{p(R \supset H)} = \frac{1/4}{3/4} = 1/3.$$

So Judy should, indeed, become less confident of R, just as van Fraassen and his co-authors claimed.[9] Extrapolating from this, it's not hard to see how similar reasoning can be applied to the original version of *Judy Benjamin*. As Eva et al. (2019) put it, for example, when your credence in an indicative conditional increases, "the antecedent becomes more informative (and hence more easily falsifiable) and less probable" (p. 468). So, even in the non-extreme version of *Judy Benjamin*, there's reason to think that Judy should become less confident of R after hearing the Captain's utterance.[10]

Surprisingly, a number of authors in the literature endorse this conclusion, despite the fact that few of them think that indicative conditionals and material conditions are equivalent. For example, alongside J. Joyce (2004), Bovens (2009), and van Fraassen et al. (1986), whom we've already mentioned, similar claims have been made by Bovens and Ferreira (2010), Eva et al. (2019), and Vasudevan (2020).

Still, this response faces serious challenges. Most obviously, there's the fact—already mentioned—that indicative conditionals are generally agreed not to be equivalent to material conditionals. This makes it something of a mystery why the learnability conditions for these conditionals should be the same. And as yet, authors who endorse this kind of response to *Judy Benjamin* haven't done much to explain why that should be.[11]

Additionally, even if we grant that conditionalizing on a material conditional gives the right results in cases where you learn an indicative conditional *with certainty*, it's not obvious how to generalize this view to uncertain learning situations. In fact, proponents of the "material conditionalization" view generally have to defend quite a different picture of updating in situations of the latter kind (see, e.g., Eva et al., 2019). This, I think, is theoretically unsatisfying. After all, conditionalization is just a special case of Jeffrey conditionalization, as we saw in my introduction. So, it's a bit strange that conditionalization applies straightforwardly to cases where you learn an indicative conditional with certainty, but Jeffrey conditionalization doesn't apply to uncertain variants thereof.

A final—and rather flat-footed—worry about this kind of response is that it just doesn't seem right to say that, when you learn an indicative conditional, your credence in the antecedent should generally go down. Indeed, one way to read *Judy Benjamin* is as a counterexample to this claim. And we can make this

[8] While this is widely agreed, it's not *universally* agreed. Important exceptions here include Lewis (1976), Jackson (1977), Grice (1989), and T. Williamson (2020). See Edgington (2021) and Rothschild (2021) for important criticisms of this view, and of Williamson's arguments for it in particular. In what follows, I'll mostly assume the falsity of the material conditional view of indicative conditionals, following the orthodoxy.

[9] More generally, a result proved by Popper and Miller (1983) shows that, if $q(-) = p(- \mid A \supset C)$, for some arbitrary $A \supset C$, then we'll usually have $q(A) < p(A)$. In other words, conditionalizing on a material conditional usually has the effect of decreasing your credence in the antecedent. The lone exception occurs when A entails C. In that case, $q(A) = p(A)$.

[10] See van Fraassen et al. (1986) for a more formal version of this argument. See also J. Joyce (2004).

[11] Important exceptions here include the views of Goldstein and Santorio (2021), Khoo (2022), and Santorio (2022). These authors attempt to provide an explanation for why learning an indicative conditional is equivalent to learning a material conditional, based on the particular semantics they provide for the former kind of conditional. Snow Zhang and I discuss this issue in greater depth in McNamara and Zhang (MS), where we also raise some concerns about it. Also, none of these authors discuss *Judy Benjamin* directly.

point even stronger. Consider, for example, yet another version of *Judy Benjamin*. This time, imagine that Judy starts out not very confident that she's in Red Territory. But after wandering around for a while, she spots a flag in the distance, which she suspects indicates that she's in Headquarters Company Territory. Her Captain then appears on the radio, and says the following:

(2) If you're in Red Territory, then you're in Headquarters Company Territory.

In this case, what should Judy's credence in R be after hearing the Captain's utterance?

Here, almost everyone I ask says that Judy's credence in R should go *up* after hearing the Captain say (2). In other words, she should become *more* confident that she's in Red Territory, after hearing the Captain's pronouncement. Notice, however, that if this is right, then we have another counterexample to the claim that learning an indicative conditional should, in general, cause your credence in the antecedent to decrease. Instead, exactly how your credence in the antecedent changes seems like it depends on specific features of the case in question.

Thus, there are, I think, good reasons to doubt that extant responses to *Judy Benjamin* of the first kind I mentioned ultimately succeed. That said, my remarks here aren't meant to be decisive, and I'm offering them only to motivate the alternative response to *Judy Benjamin* that I give in later sections. Rather than saying anything more about this kind of response, then, let's now turn to responses of the second kind, according to which we should supplement standard Bayesianism. The best-known response falling into this category is given by Douven and Romeijn (2011) and Douven (2012), drawing on ideas from Richard Bradley (2005).[12] As these authors say, cases like *Judy Benjamin* seem like they require you to change your credences according to a new rule, which Bradley (2005) calls *Adams conditionalization*:[13]

> **Adams Conditionalization**. After learning the indicative conditional $A \to C$ with certainty, your new credence in any proposition should be:
>
> $$q(-) = p(- \mid A \wedge C) \cdot p(A) + p(\neg A \wedge -). \tag{A-Cond}$$

This version of Adams conditionalization applies in cases where the learned indicative conditional seems to impose the constraint $q(C \mid A) = 1$ on your conditional credences. There's also a more general version of this rule, which applies to uncertain learning situations:[14]

> **Adams Conditionalization**. Suppose that your credences in the elements of the partition $\{A \to C, A \to \neg C\}$ shift, so that $q(A \to C) \neq p(A \to C)$.[15] Then, your new credence in any proposition should be:
>
> $$q(-) = p(A \wedge C \wedge -) \cdot \frac{q(C \mid A)}{p(C \mid A)} + p(A \wedge C \wedge \neg -) + \frac{q(\neg C \mid A)}{p(\neg C \mid A)} + p(\neg A \wedge -). \tag{A-Cond}$$

[12] Bradley (2017) also makes this response.

[13] Actually, Bradley (2005) calls this rule 'Adams conditioning'. But nothing hangs on the difference between his choice of terminology and mine. Note also that Bradley names this rule in honor of Ernest Adams, who did important early work on the relationship between conditionals and probability. See, in particular, Adams (1975).

[14] See Douven and Romeijn (2011) for an even more general version of this rule, which applies to partitions of the form $\{A \to C_i\}$. Also, see the next footnote for a brief comment on why it's legitimate to consider sets like these to be partitions in the first place.

[15] In claiming that $\{A \to C, A \to \neg C\}$ is a partition, I'm endorsing the controversial principle known as *conditional excluded middle* (CEM). According to CEM, the following is a logical truth: $(A \to C) \vee (A \to \neg C)$. Like I said, this principle is controversial (see, e.g., Lewis (1973b) for famous criticisms of it). However, the semantics for indicative conditionals I adopt in the next section validates CEM. So I won't get caught up in arguments over its validity here. Besides, as Bacon (2015) points out, CEM is a lot less controversial in the case of indicative conditionals than it is in the case of counterfactuals. And I'm only concerned with the former kind of conditional in this paper.

Now, unlike (Jeffrey) conditionalization, Adams conditionalization gives the intuitively correct results in *Judy Benjamin*, in the sense that it satisfies all of van Fraassen's desiderata. Moreover, despite how intimidating the second version of the rule might look on a first pass, Adams conditionalization turns out to have a number of intuitive properties. As Bradley (2005) points out, for example, this rule is in some sense the precise converse of (Jeffrey) conditionalization. In particular, updates by (Jeffrey) conditionalization have a property called *rigidity*, which means that, after updating, your conditional credences stay the same, even though your unconditional credences alter.[16] In the case of Adams conditionalization, in contrast, it's your conditional credences that change, while your unconditional credences in relevant propositions stay fixed. In a real sense, then, Adams conditionalization and (Jeffrey) conditionalization are like two sides of the same coin.

I'll have more to say about Adams conditionalization in §4.5. For the moment, let me just remark on this rule as a proposed solution to the *Judy Benjamin* problem. Unfortunately, while Adams conditionalization satisfies all van Fraassen's desiderata—as the reader can easily check—it still leaves a number of questions unanswered. For example, one of them is why, even though we're thinking of $A \to C$ as a proposition, this rule applies to conditionals, while ordinary (Jeffrey) conditionalization does not. Once again, this disjointedness is theoretically unsatisfying. After all, if indicative conditionals really are propositions, then we'd expect one and the same rule to apply to them as to ordinary "factual" propositions. However, proponents of Adams conditionalization offer no explanation for why this rule should apply to conditionals alone—other than the fact that it seems to get things right in a range of cases.

Moreover, while Adams conditionalization gets the right results in *Judy Benjamin*, there are some cases in which it, too, seems to get things wrong. One of them is the alternative version of *Judy Benjamin* that we looked at, where it seems like Judy's credence in R should go up, rather than stay the same. In this case, Adams conditionalization gets the wrong answer: it says that Judy's credence in R shouldn't change. In fact, this is one of the defining features of Adams conditionalization, as Bradley (2005) makes clear: in general, this update rule leaves your credence in the antecedent of an indicative conditional unaltered—which doesn't always seem like the right response.

Interestingly, even proponents of Adams conditionalization recognize this shortcoming. For example, Douven and Romeijn (2011) say that: "We are inclinded to think that Adams [conditionalization]… covers most of the cases of learning a conditional. Unfortunately, however, it would be wrong to think it covers all of them, as [the foregoing example] already shows" (p. 654). What we'd like, then, is some explanation for the distinction between the cases in which Adams conditionalization applies, and the cases in which it doesn't. But once more, no explanation has been given.

To repeat: I don't intend these remarks to constitute a decisive refutation of the second kind of response to *Judy Benjamin* that we've seen. But I do think they cast doubt on the claim that simply positing Adams conditionalization is a fully satisfactory response to that problem. Thus, what I'd like to do now is begin sketching the alternative response to this problem that I favor. As we'll see, this response will take a bit of set-up. But the reward for sticking with it is a unified solution, which accommodates all of the data we've just encountered.

4.3 Stalnaker's Thesis, Triviality, and the Sequence Semantics

Let me begin with what will seem like a bit of a detour.

[16] More precisely, if you update by conditionalization on A, then $q(- \mid A) = p(- \mid A)$. To see this, observe:

$$q(- \mid A) = \frac{q(- \wedge A)}{q(A)} = \frac{p(- \wedge A \mid A)}{p(A \mid A)} = \frac{p(- \wedge A \wedge A)/p(A)}{p(A \wedge A)/p(A)} = \frac{p(- \wedge A)}{p(A)} = p(- \mid A).$$

Something similar holds in the case of Jeffrey conditionalization. The proof there is similar.

Several authors have pointed out an apparent connection between van Fraassen's desiderata (J1)–(J3) and the notorious thesis known as *Stalnaker's thesis*.[17] The latter—which was introduced by Stalnaker (1970)[18]—posits a relationship between your credences in indicative conditionals and your conditional credences. Specifically, it says that your credence in the indicative conditional $A \to C$ should be equal to your conditional credence in C, given that A. In symbols, this is:

$$p(A \to C) = p(C \mid A). \qquad \text{(Stalnaker's Thesis)}$$

Versions of this thesis have been widely defended in the literature.[19]

To see why Stalnaker's thesis is plausible, consider an intuitive example. Suppose I'm about to roll a fair, six-sided die, when I say to you the following:

(3) If the die doesn't land on 1, then it will land on 2.

How confident should you be of the truth of this sentence? Here, most people say that your credence should be $1/5$. And assuming you give equal credence to each possible outcome of the die roll, this is just your conditional credence in 2, given that the die doesn't land on 1—which is what Stalnaker's thesis requires.

It isn't hard to see some of the ways in which this thesis is relevant to the intuitions in the *Judy Benjamin* problem. Most obviously, consider again van Fraassen's first desideratum, (J1). As Eva et al. (2019) say, this desideratum seems to be "justified by the influential idea, commonly referred to as ['Stalnaker's thesis'], that the probability of the indicative conditional 'If A, then C' is given by the corresponding conditional probability $p(C \mid A)$" (p. 464, with trivial changes of notation).[20] In particular, it seems extremely plausible that Judy's posterior credences should satisfy $q(R \to H) = 3/4$, after she's heard the Captain say (1). (And in fact, this is an assumption I'll make throughout this paper.) But then, if she also satisfies the desideratum (J1), this just *is* an instance of Stalnaker's thesis. So there's a clear connection between that thesis and van Fraassen's first desideratum in *Judy Benjamin*.

This connection is easy to spot. But there are several other ways in which Stalnaker's thesis is relevant to the *Judy Benjamin* problem—ones which I don't think have been recognized in the literature before. For example, consider again the second desideratum, namely (J2). Recall that this desideratum says that Judy's credence in R shouldn't change after hearing the Captain's pronouncement. Formally, this requirement can be cashed out by saying that Judy's credence in R is *independent* of her credence in $R \to H$. In symbols, this is written:

$$p(R \to H) = p(R \to H \mid R).$$

This is an instance of a principle that's sometimes known as *Antecedent Independence* in the literature. And it's widely recognized that this principle, together with a few plausible assumptions about the semantics of indicative conditionals, *entails* Stalnaker's thesis.[21] In fact, the same thing holds in the opposite

[17] For example, see Douven and Dietz (2011), Günther (2018), and Eva et al. (2019).

[18] Stalnaker's thesis is a precisification of the famous *Ramsey test hypothesis*, given by Ramsey (1929). It also sometimes goes by other names in the literature—e.g., 'Adams' thesis', 'the equation', or just 'the thesis'. For what it's worth, I think 'Adams' thesis' is a misnomer, since Adams himself only defended the idea that the *assertibility* of an indicative conditional goes by its conditional probability (see, e.g., Adams, 1966, 1975).

[19] Among many others, see Stalnaker (1970), McGee (1989), Bradley (2012), Bacon (2015), Khoo and Mandelkern (2018), Goldstein and Santorio (2021), Fusco (2022), Khoo (2022), Schultheis (2023), Mandelkern (forthcoming).

[20] Here, Eva et al. use the term 'Adams' thesis', rather than 'Stalnaker's thesis'. See fn. 18 for more on the distinction between these theses. For what it's worth, I think it's really Stalnaker's thesis that Eva et al. have in mind here, since they mention the *probability* of the conditional 'If A, then C', rather than its assertibility. Moreover, the phrase 'the probability of "If A, then C" really only makes sense if we think indicative conditionals as having truth-values. (After all, the probability of a proposition is its probability of *truth*.)

[21] The "plausible assumptions" I have in mind here are known as *strong centering* and *weak centering*. The first says that

direction: whenever Stalnaker's thesis is satisfied—and the relevant semantic assumptions hold—then so too does Antecedent Independence. Thus, it seems like there's a connection between Stalnaker's thesis and desideratum (J2) as well. In particular, if Judy's credences satisfy this independence condition—and if indicative conditionals obey those semantic assumptions—then she *must* satisfy Stalnaker's thesis.

These connections help to explain, I think, why van Fraassen's desiderata are so hard to accommodate within a standard Bayesian framework. After all, it's widely acknowledged that Stalnaker's thesis is itself very hard to accommodate within such a framework. To illustrate this, consider again the sentence (3), which we looked at just above. As I said there, it seems like your credence in that sentence should be $1/5$, which is what Stalnaker's thesis requires. But suppose we model this situation with a set of six, equiprobable worlds, $\mathcal{W} = \{w_1, ..., w_6\}$, where each world w_i is a world where the die lands on i (for $i = 1, ..., 6$). Then—once again—your conditional credence is $p(6 \mid \neg 1) = 1/5$. But there can be no single proposition here—i.e., no subset of the worlds $w_1, ..., w_6$—whose credence is equal to $1/5$. Instead, any such proposition must get credence equal to some multiple of $1/6$. So it looks like Stalnaker's thesis must be wrong.

This problem—which Alan Hájek (2012) calls the *wallflower problem* for Stalnaker's thesis—is closely related to the famous *triviality results* for that thesis proved by Lewis (1976) and others.[22] In rough terms, these results show that, given apparently mild background assumptions, Stalnaker's thesis can hold only in "trivial" cases—e.g., cases in which a conditional's antecedent and consequent are probabilistically independent. Lewis himself took this problem to be a reason to abandon Stalnaker's thesis. And many other philosophers have done the same. Moreover, given the apparent connection between *Judy Benjamin* and Stalnaker's thesis, we might take these results to be a reason to abandon van Fraassen's desiderata, too (as some authors have argued).

In the recent literature, however, there's been quite a lot of pushback on the triviality results for Stalnaker's thesis.[23] A number of authors have shown, for example, that if we adopt a particular *semantics* for indicative conditionals—namely, a sophisticated *sequence semantics*—then Stalnaker's thesis needn't be subject to triviality results after all. Instead, we can prove a *tenability result* for that thesis, which shows that there are non-trivial models in which it holds. Thus, the lesson of the triviality results, according to these authors, isn't that Stalnaker's thesis is implausible; it's just that we were thinking about the semantics of indicative conditionals incorrectly. In particular, if indicative conditionals obey the sequence semantics—rather than some other, more familiar semantics—then the triviality results of Lewis and Hájek need no longer hold.

In the rest of this section, I'm going to set out the sequence semantics for indicative conditionals in detail. And I'm going to show how it can be used to vindicate Stalnaker's thesis. Later on, I'll show how this semantics can be used to get the right results in *Judy Benjamin*, too. Before that, however, there's one interesting thing about the semanitcs that I want to acknowledge. This is that it's strongly inspired by van Fraassen's own work (see especially his 1976). More precisely, a version of this semantics was

---

$A \wedge C$ entails the indicative conditional $A \rightarrow C$. And the second says that indicative conditionals entail corresponding material conditionals, i.e., $A \rightarrow C$ entails $A \supset C$. Together, these principles entail the following principle about the probabilities of indicative conditionals, known as *probabilistic centering*: $p(A \wedge (A \rightarrow C)) = p(A \wedge C)$. However, probabilistic centering, together with the independence of A and $A \rightarrow C$, entails Stalnaker's thesis. Observe:

$$p(A \rightarrow C) = p(A \rightarrow C \mid A) = \frac{p(A \wedge (A \rightarrow C))}{p(A)} = \frac{p(A \wedge C)}{p(A)} = p(C \mid A).$$

More strongly, any two of Stalnaker's thesis, probabilistic centering, and the independence condition jointly entail the third. This fact will be important later on. But in the meantime, see Khoo and Santorio (2018) for further helpful discussion.

[22] As well as Lewis's paper, see, e.g., Hájek and Hall (1994), Bradley (2000), Fitelson (2015), and Goldstein and Santorio (2021). Hájek's result was originally proved in his 1989.

[23] Among others, see Bacon (2015), Khoo and Mandelkern (2018), Khoo and Santorio, 2018, Goldstein and Santorio (2021), Fusco (2022), Khoo (2022), Schultheis (2023), and Mandelkern (forthcoming).

first described by van Fraassen himself, drawing on ideas from Stalnaker (1968). This is especially notable, because van Fraassen's express purpose in developing this semantics was to get around the triviality results for Stalnaker's thesis, just as more recent authors have also done. Thus, it's striking, given the parallels between that thesis and the intuitions motivating (J1)–(J3), that no one has yet noticed the applicability of van Fraassen's semantics to the *Judy Benjamin* problem. Additionally, it's a testament to the breadth of van Fraassen's contributions that he both raised that problem, and gave us the tools required to solve it.

4.3.1 Sequence Semantics

The sequence semantics that I'll make use of throughout the rest of this paper starts with a familiar idea. This is that the truth-conditions for indicative conditionals depend on a relation of *closeness* between possible worlds. Stalnaker himself (1968) famously used this idea to develop his own semantics for indicative conditionals. Roughly speaking, his view is that a conditional $A \rightarrow C$ is true at a world w just in case the closest A-world to w is a C-world. Stalnaker then attempted to capture this idea using a device called a *selection function*—a function from propositions and worlds to possible worlds. In the sequence semantics, however, we spell things out in a slightly different way.[24]

To see how, let's start by supposing that we have a set, $\mathcal{W}$, of possible worlds, which consists of all the worlds that count as "live options" in a context (Stalnaker, 1974). For simplicity, I'll assume throughout that $\mathcal{W}$ is finite. And I'll assume that each world $w \in \mathcal{W}$ is epistemically possible for you, in the sense that you give w positive credence. I'll also assume that context supplies an *accessibility relation*, R, between the worlds in $\mathcal{W}$, which represents something like the possibilities left open by your evidence at w. So, R says that v is accessible from w, written '$w\mathsf{R}v$', just in case v is compatible with your evidence at w. I'll assume this relation is reflexive. And I'll write '$\mathsf{R}(w)$' for the set of worlds accessible from w. Finally, I'll often assume that R is an equivalence relation on $\mathcal{W}$, which relates every world to every other world. And when that's the case, follow Mandelkern (forthcoming) and say that the context is *transparent*. (As we'll see, however, not every context is a transparent context. This will be important in §4.5.)

Now, a *sequence* of worlds is n-tuple of the worlds in $\mathcal{W}$, without repetitions. For example, if $\mathcal{W} = \{w_1, w_2, w_3\}$, then one sequence we can generate from this set is $\langle w_1, w_2, w_3 \rangle$. Interpretationally, sequences represent how *close* worlds are to the first world in the sequence. So, $\langle w_1, w_2, w_3 \rangle$ says, for instance, that w_1 is the closest world to itself, w_2 is the next closest, and so on. In what follows, I'll assume that sequences consist of all and only the worlds that are accessible from the first world in the sequence. So, if the sequence $\langle w_1, w_2, w_3 \rangle$ is a legitimate such sequence, then this implies that $\mathsf{R}(w_1) = \mathcal{W}$. Additionally, I'll assume that any such sequence is a legitimate sequence in the context, in the sense that is supplies an appropriate closeness ordering in the context. I'll call sequences like this *admissible* sequences.

Now, in the sequence semantics, sequences function as the *points of evaluation* for indicative conditionals. That is, a conditional sentence is true or false *at a sequence of worlds*, rather than at a possible world (as is more usually the case). To make this precise, let me introduce some useful notation. Going forward, I'll write 'c' to denote a context, 's' to denote a sequence, and 's_A' to denote the first A-world in s. Then, with this notation in hand, we can state the sequence semantics for indicative conditionals as follows:

Sequence Semantics. $[\![\text{If } A, \text{ then } C]\!]^{s,c} = 1$ if and only if $s_A \in C$.

Informally, this says that a sentence 'If A, then C' is true at a sequence s, in a given context, just in case the first A-world in that sequence is a C-world.

[24] Incidentally, if selection functions obey a handful of natural constraints, then every pair consisting of a world and a selection function determines a sequence of possible worlds, and vice versa. So, in a sense, we could alternatively work with Stalnaker's selection functions, rather than with sequences. As we'll see, however, it's a bit easier to work directly with sequences, mathematically speaking. But in essence, the sequence semantics I give below is closely related to Stalnaker's semantics. See Dorr and Mandelkern (MS) for more on this relationship.

There are a couple of things to note about this semantics. The first is that it bears a strong similarity to Stalnaker's semantics. That latter semantics, recall, said that 'If A, then C' is true at a world w just in case the closest A-world to w is a C-world. If sequences thus function as closeness orderings of worlds, then the sequence semantics says much the same. The main difference is just that, rather than thinking of there being a single, *true* closeness ordering with which to interpret indicative conditionals, we instead allow that context can supply a *range* of admissible orderings.

The main upshot of this is that, on the sequence semantics for indicative conditionals, indicative conditionals will often be *indeterminate* at possible worlds. Instead, their truth-values depend on more fine-grained possibilities—namely, on relations of closeness *between* possible worlds. To see why this view is plausible, consider again the sentence (3):

(3) If the die doesn't land on 1, then it will land on 2.

Once more, suppose we model this situation with six worlds, $w_1, ..., w_6$. And suppose you're ignorant about how the die will actually land. Then, ask yourself: What's the truth-value of (3) at the world where the antecedent is false—namely, w_1? Intuitively, it's very hard to answer that question—and the sequence semantics tells us why. Roughly speaking, the reason is that the coarse-grained, "descriptive" facts that obtain at w_1—i.e., that the die landed on 1 at that world—don't give us any obvious way of settling how the die landed *if* it landed on some number other number. For that, we need to say what's the *closest* world to w_1 at which the antecedent of (3) is true. And that's the role that sequences play in the present construction.

Later on, we'll see that this "fine-grained" way of thinking about indicative conditionals is really the key to getting around the triviality results for Stalnaker's thesis, given by Lewis and others. It also turns out to be the key to satisfying van Fraassen's desiderata (J1)–(J3) in *Judy Benjamin*. In the meantime, however, let me say one more thing by way of motivation for this semantics. This is that it fits very naturally with a *contextualist* view about indicative conditionals—i.e., a view according to which the proposition expressed by an indicative conditional sentence depends on the context in which it's uttered.

To see why this pairing is so natural, first recall how I glossed the set of worlds $\mathcal{W}$ above. There, I said that we should think of this set as the set of worlds that count as "live options" in the context. If, however, we think of sequences as consisting of just these worlds, then it looks like the sequences that we use to evaluate indicative conditionals can change depending on the context. In particular, in different contexts, different possible worlds will count as "live options". And for that reason, different sequences will be used to evaluate indicative conditionals, too.

The other—and more important—reason for thinking that the van Fraassen-style semantics is a contextualist semantics, however, is that, for philosophers who endorse contextualism, there isn't just one privileged notion of closeness that's suitable across all contexts. Instead, exactly what we *mean* by 'closeness' can change depending on the context. In §4.5, we'll look at examples to see that this is so. But for now, just note that this falls out very naturally of the present construction. After all, recall that I said that the "admissible" sequences in a context are constrained by the accessibility relation. That is, a sequence $\langle w_1, w_2, ..., w_n \rangle$ that we use to evaluate indicative conditionals should consist only of worlds that are accessible from the first world, according to the contextually specified accessibility relation R. Thus, in different contexts where different accessibility relations are in play, different notions of closeness will count as admissible. Thus, there are really two different ways in which the sequence semantics is naturally interpreted as a contextualist semantics. You needn't interpret the semantics in this way, but I think it's very plausible to do so. And besides, van Fraassen himself sees the sequence semantics as a contextualist semantics, just as most recent authors working with this semantics have also done.

4.3.2 Credences

With the sequence semantics for indicative conditionals now in place, let's turn to a different issue. On a first pass, you might think this semantics raises more problems than it solves. One obvious concern is that, as I've spelled things out here, indicative conditional propositions need no longer correspond to sets of worlds. But earlier, I assumed that propositions *are* sets of worlds, and thus that your credence function, p, is defined only over such sets. How, then, are you supposed to assign credences to indicative conditionals in a meaningful way, in this new, sequence-based setting?

In order for you to do this, we'll need to find a way of *extending* your credence function, so that it's defined over sequences, and not just over worlds. To make this extension, then, I'll once again take inspiration from van Fraassen (1976). Specifically, following a suggestion of Goldstein and Santorio (2021) and Khoo (2022)—who in turn draw on van Fraassen—we'll imagine that your credences in ordinary "factual" propositions are represented by a credence function, p^-, defined only over worlds. Then, given this function, we'll extend p^- to your full credence function, p, over sequences, using a recursive procedure (below, I write '$[w]$' for the set of sequences beginnning with w, and '$[w_1, ..., w_k]$' for the set of sequences that share the same k-length initial segment, namely $w_1, ..., w_k$, in that order):

(i) $p([w]) = p^-(w)$,

(ii) $p([w_1, ..., w_k]) = p([w_1, ..., w_{k-1}]) \cdot p^-(w_k \mid \mathsf{R}(w) - \{w_1, ..., w_{k-1}\})$.

Heuristically, we can think of this recursive procedure as saying that your credence in a sequence, $\langle w_1, w_2, ..., w_n \rangle$, is your credence that you'd draw those world from an urn, in that order, and without replacement. This is a very natural way of extending p^- to a credence function, p, defined over sequences. Roughly, it says that the credences you assign to sequences are "parasitic" on the credences you assign to worlds, in the sense that your credences about relations of closeness derive from your credences about ordinary facts. Whenever p comes from a function p^- in this way, I'll say that p is *well-behaved*.[25]

Now, it's easy to see that defining p according to (i) and (ii) preserves the credences that p^- assigns to "factual" propositions. To quickly illustrate this anyway, however, consider again the toy example, where $\mathcal{W} = \{w_1, w_2, w_3\}$. Suppose that $p^-(w_1) = 1/2$, $p^-(w_2) = 1/3$, and $p^-(w_3) = 1/6$. Imagine that A is a factual proposition true at w_1 and w_2, and C is a factual proposition true at w_2 and w_3. Then, it follows that $p^-(A) = 5/6$, and $p^-(C) = 1/2$. Moreover, assuming the context is transparent, we have that:

$$\begin{aligned} p(\langle w_1, w_2, w_3 \rangle) &= p([w_1, w_2]) \cdot p^-(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= p([w_1]) \cdot p^-(w_2 \mid \mathcal{W} - \{w_1\}) \cdot p^-(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= p^-(w_1) \cdot p^-(w_2 \mid \mathcal{W} - \{w_1\}) \cdot p^-(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= 1/2 \cdot 2/3 \cdot 1 \\ &= 1/3. \end{aligned}$$

Similar calculations show that $p(\langle w_1, w_3, w_2 \rangle) = 1/6$, $p(\langle w_2, w_1, w_3 \rangle) = 1/4$, and $p(\langle w_2, w_3, w_1 \rangle) = 1/12$. And taking the sum of your credences in all of these sequences gives $p(A) = 5/6$, as desired. It's not hard to show that something similar holds for the proposition C .

This feature of the recursive procedure has an especially important upshot, at least when it comes to the issues that we're interested in here. To see what it is, first notice that, because p preserves the credences that p^- assigns to factual propositions, it preserves the *conditional* credences that p^- assigns as well. (After all, conditional credences are just ratios of unconditional credences—take another look at the definition of conditionalization, in §4.1, to see this.) For example, consider the conditional credence $p^-(C \mid A)$. This is

[25] For alternative proposals about how to extend your credence function to a function defined over sequences, see van Fraassen (1976) or Mandelkern (forthcoming).

equal to $2/5$. And a few back-of-the-envelope calculations show that this is the credence that p assigns to C given A as well.

Strikingly, however, p also assigns this credence to the indicative conditional $A \to C$, at least given our assumption that the context is transparent. To see this, just note that $A \to C$ is true at three sequences out of the six sequences we can construct from the worlds in $\mathcal{W}$, namely: $\langle w_2, w_1, w_3\rangle$, $\langle w_2, w_3, w_1\rangle$, and $\langle w_3, w_2, w_1\rangle$. Once again, a few tedious calculations show that $p(\langle w_2, w_1, w_3\rangle) = 1/4$, $p(\langle w_2, w_3, w_1\rangle) = 1/12$, and $p(\langle w_3, w_2, w_1\rangle) = 1/15$. And taking the sum of your credence in all these sequences gives $p(A \to C) = 2/5$.[26] So, what we have here is a case in which Stalnaker's thesis (and note that it's satisfied non-trivially, since $p(A)$ and $p(C)$ are probabilistically independent). Contrary to what authors like Lewis and Hájek claimed, then, Stalnaker's thesis can hold non-trivially after all.

As it turns out, this "tenability result" for Stalnaker's thesis holds quite widely. Specifically, if p is a well-behaved credence function and the context is transparent, then we get the following important result:[27]

Theorem 2 (van Fraassen, 1976; Goldstein and Santorio, 2021; Khoo, 2022)**.** *Let p be a well-behaved credence function. Let the context be transparent. Then, for all factual propositions A and C:*

$$p(A \to C) = p(C \mid A).$$

That is, Stalnaker's thesis holds.[28]

This result was first proved by van Fraassen (1976), using a slightly different set-up. Essentially, it establishes the *tenability* of Stalnaker's thesis, given the sequence semantics. Because that semantics appeals to more fine-grained possibilities than just sets of possible worlds, results like Hájek's wallflower result no longer apply. And as we'll see in a moment, it's really this result that allows us to get the right answers in the *Judy Benjamin* problem, too.

4.3.3 Updating

There's one last piece of the puzzle I need to put in place, before we can return to *Judy Benjamin*. This is to say how your credences should *change* in this new, sequence-based setting. In standard Bayesianism, changes in credence are governed by conditionalization or Jeffrey conditionalization, depending on the circumstances. And in my view, the same rules still apply. The only difference now is that, whereas before we were thinking of changes in credence as applying (only) to sets of worlds, we'll now think of them as applying to arbitrary sets of sequences.

For completeness, then, let me re-state the Bayesian update rules, this time in the special cases in which they apply to indicative conditionals. The first rule is:

[26] In the first case, for instance:

$$\begin{aligned} p(\langle w_2, w_1, w_3\rangle) &= p([w_2, w_1]) \cdot p^-(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= p([w_2]) \cdot p^-(w_1 \mid \mathcal{W} - \{w_2\}) \cdot p^-(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= p^-(w_2) \cdot p^-(w_1 \mid \mathcal{W} - \{w_1\}) \cdot p^-(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= 1/4. \end{aligned}$$

The other cases are computed similarly.

[27] It's possible to extend this result even further, to other kinds of contexts. Unfortunately, however, I've not been able to show how to do that here.

[28] Or rather, a *restricted* version of Stalnaker's thesis holds, because, as I've stated it, the thesis applies only when A and C are factual. In other words, it doesn't say whether Stalnaker's thesis is satisfied in cases where A or C are themselves indicative conditionals (or modals). It's possible to extend the result stated here so that Stalnaker's thesis holds unrestrictedly—see Bacon (2015) for how to do this. However, doing so does involve jettisoning the interpretation of sequences as encoding relations of *closeness*. Moreover, when it comes to embedded conditionals, intuitions about the validity of Stalnaker's thesis are mixed. So I won't say anything more about unrestricted versions of the thesis.

Conditionalization. After learning $A \to C$ with certainty, your new credence in any proposition should be equal to your old conditional credence in that proposition, given $A \to C$. Formally:

$$q(-) = p(- \mid A \to C).$$

Similarly, in the case of Jeffrey conditionalization, we have:

Jeffrey Conditionalization. Suppose your credences in the elements of the partition $\{A \to C, A \to \neg C\}$ shift so that $q(A \to C) \neq p(A \to C)$. Then, your new credence in any proposition should be:

$$q(-) = p(- \mid A \to C) \cdot q(A \to C) + p(- \mid A \to \neg C) \cdot q(A \to \neg C).$$

Once again, these are more-or-less the same rules for updating that I stated at the outset. The only difference now is that we're allowing the *content* of the relevant propositions to be different from what we initially assumed.

That, in a nutshell, is my whole theory of updating. Before we move on, however, I want to make a few remarks about this theory. In particular, I want to address a potential worry you might have about it. You might be concerned, for example, that this theory can't be right, given my assumption that your credences in sequences are "parasitic" on your credences in worlds.[29]

To see the issue I'm getting at, let's consider an example. Think again about the toy case, where $\mathcal{W} = \{w_1, w_2, w_3\}$. Once more, suppose that $p^-(w_1) = 1/2$, $p^-(w_2) = 1/3$, and $p^-(w_3) = 1/6$. Let $A = \{w_1, w_2\}$ and $C = \{w_2, w_3\}$. Suppose that the context is transparent. And finally, suppose that p is an extension of p^- that's well-behaved. Then, as we saw before, the credence that p assigns to the indicative conditional $A \to C$ is $2/5$. But suppose now that you have a learning experience, which causes you to become certain of $A \to C$, so that $q(A \to C) = 1$. Then, won't this result in the function p no longer being well-defined, since some worlds in $\mathcal{W}$ will now get probability 0? In particular, $A \to C$ is false at all the sequences beginning with w_1 (as the reader can easily check). So, conditionalization implies that your credence in w_1 should now be 0. However, the recursive procedure (i)–(ii) requires that you assign positive credence to each world $w \in \mathcal{W}$. So how are we to square this fact with my theory of updating?

The answer here lies in my assumption that the sequences we use to evaluate indicative conditionals should consist only of the worlds that count as "live options" in the context. Thus, once you assign credence 0 to w_1 in the foregoing example, that world no longer counts as a live option, and so the set of admissible sequences in the context will change. In particular, once you've conditionalized on the proposition $A \to C$, the only sequences left over will be $\langle w_2, w_3 \rangle$, $\langle w_3, w_2 \rangle$, since these are the only sequences we can construct from wolrds with positive probability. Then, your new credences in these sequences should be given in accordance with the new credences you assign to the worlds which compose them.

There's more to be said about this, but I'll defer further discussion for §4.5. There, we'll see that the theory of updating I've sketched in this section turns out to have a surprising consequence, especially in cases like the one I've just described. For now, however, let's return to *Judy Benjamin*. What I want to show is that, if Judy updates in the way I've just specified—particularly, if she updates by Jeffrey conditionalization—then all of van Fraassen's desiderata are satisfied. And that's surprising since van Fraassen's original claim was that Judy can't satisfy (J1)–(J3) if she updates in a Bayesian fashion.

4.4 *Judy Benjamin* Redux

Recall the original version of *Judy Benjamin*. In that case, Judy and her platoon are lost in an unknown territory composed of four quadrants, equally-sized. The territory is divided into a Red Territory and a

[29] Thanks to Paolo Santorio for pushing me to consider this objection.

Blue Territory. And each of these territories is subdivided into a Headquarters Company Territory and a Second Company Territory. Judy's initial credence that she's in any one of these territories is equal to $1/4$. She then hears the Captain say the following on the radio:

(1) The probability is $3/4$ that if you're in Red Territory, then you're in Headquarters Company Territory.

The question now is: How should Judy's credences change after she hears the Captain say this? In particular, can she satisfy van Fraassen's desiderata, (J1)–(J3), if she updates by (Jeffrey) conditionalization?

To see that she can, let's focus on a simplified version of the problem in this section. (I discuss the more general version of the problem in Appendix 4.6 below. I focus on the simple version here just to make the calculations easy.) Specifically, imagine that Judy believes only four possible words could be actual. Then, her set of epistemically possible worlds is just $\mathcal{W} = \{w_1, w_2, w_3, w_4\}$, and each of these worlds corresponds to the possibility that Judy is in some sub-region of the territory. More precisely, imagine that Judy's epistemic possibilities are summarized by this table:

	H	$\neg H$
R	w_1	w_2
$\neg R$	w_3	w_4

Table 4.1: Judy's Epistemic Possibilities

So, the possibility that Judy is in both Red Territory and Headquarters Company Territory is just the possibility that w_1 is actual. And the possibility that she's in Red Territory and Second Company Territory is the possibility that w_2 is actual. And so on. Now, as I said, Judy initially gives each of these possibilities equal credence. So for any world w_i, with $i = 1, ..., 4$, we have that $p^-(w_i) = 1/4$.

Now, there are a couple of substantive—but reasonable—assumptions I'm going to make in order to secure my result about updating. The first is just that Judy's credences in indicative conditionals are given by a well-behaved credence function p, which extends a credence function p^-. Given the credences she initially assigns to the worlds $w_1, ..., w_4$—and assuming the context is transparent—it's then easy to show, using the recursive procedure (i)–(ii), that Judy's credence in $R \to H$ is initially $1/2$. In other words, she's $50/50$, at the start, about whether, if she's in Red Territory, then she's in Headquarters Company Territory.

Next, the second assumption I'll make is that, after hearing the Captain say (1), Judy's credence in the indicative conditional $R \to H$ rises to $3/4$. In symbols, this is:

$$q(R \to H) = 3/4.$$

This seems like a perfectly natural thing to assume, given the way we spelled out the case. And anyway, it's made by many other authors in the literature (see, e.g., Douven and Romeijn, 2011; Douven, 2012).

Finally, the last assumption I'll make—which I made in passing above—is that the context is transparent. In other words, given Judy's set of epistemically possible worlds, $\mathcal{W} = \{w_1, w_2, w_3, w_4\}$, I'll assume that every sequence we can construct from these four worlds counts as an admissible sequence in the context.

Now, given these assumptions, we can then show that, if Judy updates by Jeffrey conditionalization on the partition $\{R \to H, R \to \neg H\}$, then all of van Fraassen's desiderata are satisfied. And this is contrary to what van Fraassen originally claimed.

In the first case, it's almost trivial to show this. To see why, recall that desideratum (J1) required that Judy's posterior credences be such that $q(H \mid R) = 3/4$. However, above we assumed that, after hearing the Captain say (1), Judy's credence in $R \to H$ rises to $3/4$. So, by Theorem 2, discussed in the last section, it follows that $q(H \mid R) = 3/4$ as well. After all, that theorem says that, in any transparent context, Stalnaker's thesis is satisfied. So desideratum must (J1) hold.

Now let's turn to the second desideratum, (J2). This case is also fairly straightforward. According to this desideratum, Judy's credences should satisfy $q(R) = p(R) = 1/2$. In other words, her credence in R shouldn't change after hearing the Captain say (1). Once again, this desideratum follows straightforwardly from Theorem 2, if Judy updates her credences by Jeffrey conditionalization. After all, in §4.3 I noted that—given some plausible semantic assumptions about indicaitive conditionals—Stalnaker's thesis implies that R and $R \to H$ are probabilistically independent. Thus, since the sequence semantics for indicative conditionals satisfies those assumptions,[30] and since Judy satisfies Stalnaker's thesis in the present context, it follows that $q(R) = p(R) = 1/2$, since her credences in R and $R \to H$ are independent. The upshot is that, even after she updates her credences in $R \to H$, Judy's credence in the antecedent, R, doesn't change. So desideratum (J2) holds.

Finally, let's turn to desideratum (J3). Recall that this desideratum said that, for any proposition, any for any $X \in \{R \wedge H, R \wedge \neg H, \neg R\}$, we should have:

$$q(- \mid X) = p(- \mid X). \tag{J3}$$

In other words, Judy's conditional credences, given one of $R \wedge H$, $R \wedge \neg H$, or $\neg R$ shouldn't change after she hears the Captain say (2). The first two cases here are easy. As I've spelled things out, $R \wedge H = w_1$. And for any "factual" proposition A, $p(A \mid w_1) \in \{0, 1\}$. In other words, conditional on the world w_1 being actual, every factual proposition either has probability 1 or 0 (since possible worlds settle the truth-values for all factual propositions). But of course, that doesn't change when Judy updates her credences to the new function q. So, we have that: $q(A \mid w_1) = p(A \mid w_1)$, for any factual proposition A. At least if we restrict ourselves to factual propositions, then, the first case of (J3) is satisfied. (And note that van Fraassen himself *does* restrict things to factual propositions.)

An exactly parallel argument establishes the second case of (J3), i.e., the case in which $q(- \mid R \wedge \neg H) = p(- \mid R \wedge \neg H)$. So we only need to check the last case to see that desideratum (J3) is satisfied. In other words, we only need to check that $q(- \mid \neg R) = p(- \mid \neg R)$, and then we're done. Unfortunately, showing that the last case holds is a little bit more involved, and requires some genuine calculations. So I'll relegate those calculations to a footnote.[31] The basic idea, however, is that the desideratum follows from the rigidity property of Jeffrey conditionalization (see §4.2). Thus, the desideratum (J3) holds in example as well, which means that all of van Fraassen's desiderata are satisfied. And note that this is so, even though we assumed Judy updates her credences by Jeffrey conditionalization.

For fans of the Bayesian update rules, this should come as good news. In effect, it shows that philosophers like Douven (2012) were wrong to say that "updating on conditionals [seems to be] very different from standard Bayesian updating" (p. 240), at least when it comes to cases like *Judy Benjamin*.

[30] In particular, note that the sequence semantics satisfies *probabilistic centering*. To see this, just note that $A \wedge (A \to C)$ in the sequence semantics is the set of all sequences beginning with A at which $A \to C$ is true. But of course, our semantics says that $A \to C$ is true at a sequence, just in case the first A-world in that sequence is a C-world. It follows immediately that $A \wedge (A \to C) = A \wedge C$, and so the probabilities of the two must be the same. This is what probabilistic centering requires.

[31] The calculations are:

$$\begin{aligned}
q(- \mid \neg R) &= \frac{q(- \wedge \neg R)}{q(\neg R)} \\
&= \frac{p(- \wedge \neg R \mid R \to H) \cdot q(R \to H) + p(- \wedge \neg R \mid R \to \neg H) \cdot q(R \to \neg H)}{p(\neg R \mid R \to H) \cdot q(R \to H) + p(\neg R \mid R \to \neg H) \cdot q(R \to \neg H)} \\
&= \frac{p(- \wedge \neg R \mid R \to H) + p(- \wedge \neg R \mid R \to \neg H)}{p(\neg R \mid R \to H) + p(\neg R \mid R \to \neg H)} \\
&= \frac{p(- \wedge \neg R)}{p(\neg R)} \\
&= p(- \mid \neg R)
\end{aligned}$$

4.5 General Results

The *Judy Benjamin* problem is probably central problem that philosophers have focused on in the literature, suggesting indicative conditionals pose a special problem for the standard Bayesian update rules. At the same time, however, it's only a single example. And it wouldn't say very much if my theory got things right in this case, but didn't have any wider upshots. What I want to do now, then, is begin closing the paper by outlining some of the more general results that can be gleaned from my theory of updating. As I said in §4.2, one of the most attractive things about this theory is that it accommodates the data that motivated the two most common types of response to *Judy Benjamin* and related examples. So let's begin by seeing how that is so.

I'll start with a bit of a warm-up. In §4.2, we heard an argument for the claim that when you learn an indicative conditional $A \rightarrow C$ with certainty, you should conditionalize on a corresponding material conditional, $A \supset C$. Roughly speaking, this argument was motivated by the idea that learning $A \rightarrow C$ with certainty seems to impose the constraint $q(C \mid A) = 1$ on your conditional credences. And once we have that, it follows by the ratio formula that $q(A \supset C) = 1$.

Now, like I said, I'm not convinced that learning an indicative conditional with certainty requires you to conditionalize on a corresponding material conditional. My chief objections to this view were that (i) indicative conditionals and materials conditionals are generally agreed not to be equivalent to one another; and (ii) the material conditionalization view implies that, when you learn an indicative conditional with certainty, your credence in the antecedent should generally go down. This latter fact, I argued, doesn't always seem like the right response (indeed, *Judy Benjamin* is arguably a counterexample to this claim).

At the same time, however, I think the material conditionalization view gets *something* importantly right. Specifically, in my view, it's correct to say that learning an indicative conditional with certainty is *qualitatively* identical to learning a material conditional, in the sense that becoming certain of the one entails becoming certain of the other. In fact, this idea turns out to be implied by the sequence semantics for indicative conditionals, to which I've been appealing. Boylan and Schultheis (2021) call this the *qualitative thesis*:

Theorem 3 (Boylan and Schultheis, 2021)**.** *Let p be a well-behaved credence function, that extends a credence function p^-. Let A and C be factual propositions. Then, $p(A \rightarrow C) = 1$ if and only if $p(A \supset C) = 1$.*

My statement of Theorem 3, the qualitative thesis, is a little different to Boylan and Schultheis's statement of it. So I've included a proof of this result in the Appendix below. In words, you can think of it as saying that, if indicative conditionals have a sequence semantics, then you're certain of $A \rightarrow C$ just in case you're certain of $A \supset C$. In other words, you're certain of an indicative conditional just in case you're certain of a corresponding material conditional.

I think this is really the strongest thing that the argument from §4.2, due to van Fraassen et al. (1986), establishes. After all, that argument said that, if learning an indicative conditional imposes the constraint $q(C \mid A) = 1$ on your conditional credences, then your credences should also satisfy $q(A \supset C) = 1$. Notice, however, that it's a further step from this to the claim that you should therefore update by conditionalizing on $A \supset C$. In other words, just because $q(A \supset C) = 1$, this need not imply that $q(-) = p(- \mid A \supset C)$. On the contrary, the theory of updating I sketched in §4.3 delivers the datum that $q(A \supset C) = 1$. But it doesn't say that how you should arrive at this posterior credence should be by conditionalizing on a material conditional.

Thus, the theory I sketched in §4.3 accommodates the chief data point that motivated the first kind of response to *Judy Benjamin* that we considered. You might also notice that, in that problem, my theory agrees with the result of Adams conditionalization—which was essentially the second kind of response to *Judy Benjamin*. That is, Adams conditionalization satisfies van Fraassen's desiderata (J1)–(J3), and so,

too, does my theory. This might lead you to wonder whether there's any sort of systematic connection between Adams conditionalization and Jeffrey conditionalization on a set of sequences.

It turns out there is. In particular, it turns out that in any transparent context, Jeffrey conditionalization on a set of sequences is equivalent to Adams conditionalization. More precisely, we have the following result, which is the main formal result of this paper:

Theorem 4. *Let p be a well-behaved credence function, and let A and C be factual propositions. Suppose that the context is transparent, and suppose that q comes from p by Jeffrey conditionalization on the partition $\{A \to C, A \to \neg C\}$. Then, for any proposition:*

$$q(-) = p(A \wedge C \wedge -) \cdot \frac{q(C \mid A)}{p(C \mid A)} + p(A \wedge C \wedge \neg -) + \frac{q(\neg C \mid A)}{p(\neg C \mid A)} + p(\neg A \wedge -).$$

That is, Jeffrey conditionalization on $\{A \to C, A \to \neg C\}$ is equivalent to Adams conditionalization.

Thus, Theorem 4 says that, in any transparent context, Jeffrey conditionalization is equivalent to the results you get by Adams conditionalization on a set of worlds. In a sense, then, Adams conditionalization was all along a form of Jeffrey conditionalization—really, it was something like Jeffrey conditionalization in disguise.

Of course, since conditionalization is just a special case of Jeffrey conditionalization, Theorem 4 implies that conditionalization on $A > C$ is equivalent to the simpler version of Adams conditionalization, too:

Corollary 1. *Let p be a well-behaved credence function, and let A and C be factual propositions. Suppose that the context is transparent, and suppose that q comes from p by conditionalization on $A \to C$. Then, for any proposition:*

$$q(-) = p(- \mid A \wedge C) \cdot p(A) + p(- \wedge \neg A).$$

That is, conditionalization on $A \to C$ is equivalent to Adams conditionalization.

This, I think, is an especially attractive result, because it helps us relate things back to Theorem 3, the qualitative thesis. To see what I mean, consider again the simplified version of *Judy Benjamin*, which we considered in the previous section. In that case, recall, Judy's epistemic possibilities were as described in the following table:

	H	$\neg H$
R	w_1	w_2
$\neg R$	w_3	w_4

Table 4.2: Judy's Epistemic Possibilities

Now, suppose that in this case, Judy hears her Captain say (2), instead of saying (1):

(2) If you're in Red Territory, then you're in Headquarters Company Territory.

Then, in my view, Judy should conditionalize on the indicative conditional $R \to H$. And if the context is transparent, the result of doing so is equivalent to the results we get by Adams conditionalization on a set of worlds (that's what Corollary 1 tells us). In particular, Adams conditionalization implies that after updating, Judy's credence in each world w_i will be as in the left table below:
But in contrast, material conditionalization implies that Judy's credence in each world w_i will be as in the table on the right. Thus, how Judy distributes her credences among the worlds $w_1, ..., w_4$ after learning $R \to H$ is different in the two cases. But qualitatively speaking, the two learning experiences are the same. In particular, in both cases we have that $q(H \mid R) = 1$, and also that $p(R \to H) = p(R \supset H) = 1$.

	H	$\neg H$
R	$1/2$	0
$\neg R$	$1/4$	$1/4$

	H	$\neg H$
R	$1/3$	0
$\neg R$	$1/3$	$1/3$

So learning the indicative conditional $R \rightarrow H$ is qualitatively equivalent to learning $R \supset H$, even though the two cases aren't probabilistically equivalent.

Finally, the theory I've laid out here is also flexible enough to explain why Adams conditionalization sometimes fails. To see this, notice that Theorem 4 says only that Jeffrey conditionalization on a set of sequences is equivalent to Adams conditionalization *in transparent contexts*. In other kinds of contexts, however, this equivalence isn't guaranteed. In particular, in contexts where a conditional $A \rightarrow C$ and its antecedent aren't probabilistically independent, my theory predicts that updating by Jeffrey conditionalization will diverge from the results of Adams conditionalization.

To illustrate this, let's go back to the version of Judy Benjamin that I considered towards the end of §4.2, where Judy wasn't very confident, to start with, that she's in the Red Territory. Recall that she then spots a flag in the distance, which she suspects could indicate that she's in Headquarters Company Territory. Then, Judy's Captain says (2) on the radio:

(2) If you're in Red Territory, then you're in Headquarters Company Territory.

As we then heard, it seems like Judy's credence in R should go up in the case, rather than go down or stay the same (which is what material conditionalization and Adams conditionalization, respectively, require).

We can accommodate this datum in my theory by noting that the extra information Judy has—namely, that there's a flag in the distance—seems to impose a natural constraint on the context's accessibility relation, R. In particular, it seems to supply the context with a natural background partition, $\{F, \neg F\}$, where F is the proposition that the flag indicates Headquarters Company Territory, and $\neg F$ is the proposition that the flag indicates Judy is not in Headquarters Company Territory. Now, imagine that Judy's credences are such that $p(R \rightarrow H \mid F)$ is high, while $p(R \rightarrow H \mid \neg F)$ is low, as seems natural. Then, learning $R \rightarrow H$ with certainty should increase credence Judy's credence in the proposition F. But also, if her credence in F increases, then she should also increase her credence in the antecedent of the conditional R. Thus, what we get in this case is a failure of the independence of $R \rightarrow H$ and R—and that's why Adams conditionalization doesn't apply.

Thus, my theory can accommodate the data that motivated the other two responses to the *Judy Benjamin* problem. In particular, it accommodates the idea that learning an indicative conditional is closely related to learning a material conditional. But we've also seen that updating by (Jeffrey) conditionalization on a set of sequences is equivalent to Adams conditionalization in certain contexts, and thus gets the right answers in *Judy Benjamin*. At the same time, however, my theory allows us to explain why Adams conditionalization sometimes *doesn't* seem like the right response to learning a conditional. In particular, when a conditional $A \rightarrow C$ and its antecedent aren't independent—as seems natural in some contexts—my theory diverges from the prescriptions of Adams conditionalization. And that seems correct. Thus, my theory is at once more flexible, and allows us to accommodate a larger quantity of data, than other responses to *Judy Benjamin* that have so far been proposed.

4.6 Conclusion

I began this paper by spelling out a problem for standard Bayesianism—namely, that this theory seems to give the wrong results in cases where you learn an indicative conditional. I argued, however, that this problem can be resolved if we adopt a particular semantics for indicative conditionals—namely, a *sequence*

semantics, originally posed by Bas van Fraassen (1976). Van Fraassen's semantics was originally introduced to get around the well-known triviality results for Stalnaker's thesis due to Lewis (1976) and others. And in the preceding, I tried to draw out a connection between those results, and van Fraassen's own *Judy Benjamin* problem. We saw that, by co-opting the tenability results for Stalnaker's thesis, which make use of van Fraassen's semantics, we get resolve the *Judy Benjamin* problem. And more broadly, we can show that standard Bayesian need not be threatened by the problem of learning 'if'.

Appendix

Calculations in *Judy Benjamin*

In this Appendix, I calculate explicitly the results discussed in §4.4. That is, I show that, if the context is transparent, then, if Judy updates by Jeffrey conditionalization on the partition $\{R \to H, R \to \neg H\}$ after hearing the Captain say (1), her posterior credences will satisfy all of van Fraassen's three desiderata.

To see this, first suppose that the context is transparent. Then, since Judy spreads her (prior) credences evenly over the propositions $R \wedge H$, $R \wedge \neg H$, etc., it follows that $p(R \to H) = 1/2$. Furthermore, by Theorem 2, it follows $p(H \mid R) = 1/2$.

Now suppose that, after hearing the Captain say (1), Judy's posterior credence in $R \to H$ rises to $3/4$. Then, again by Theorem 2, it follows that $q(H \mid R) = 3/4$. So desideratum (J1) is satisfied. (This is basically the same argument that I made in §4.4.)

Now consider (J2). Suppose that, after Judy hears the Captain's testimony, she updates her credences by Jeffrey conditionalization on the partition $\{R \to H, R \to \neg H\}$. By the ratio formula for conditional probability, her prior conditional credence $p(R \mid R \to H)$ is:

$$p(R \mid R \to H) = \frac{p(R, R \to H)}{p(R \to H)}.$$

But then, by Probabilistic Centering, it follows that $p(R, R \to H) = p(R, H) = 1/4$. Thus, plugging this into the ratio formula above gives: $p(R \mid R \to H) = 1/2$. Parallel reasoning then shows that $p(R \mid R \to \neg H) = 1/2$. So, finally, Jeffrey conditionalization implies:

$$\begin{aligned} q(R) &= p(R \mid R \to H) \cdot q(R \to H) + p(R \mid R \to \neg H) \cdot q(R \to H) \\ &= 1/2 \cdot 3/4 + 1/2 \cdot 1/4 \\ &= 1/2. \end{aligned}$$

So $q(R) = 1/2$, and desideratum (J2) is satisfied.

Now turn finally to desideratum (J3). We want to show that $q(- \mid \neg X) = p(- \mid \neg X)$, for any factual proposition, and where $X \in \{R \wedge H, R \wedge \neg H, \neg R\}$. I already showed this in the main text, in the case where $X = \neg R$. So I'll focus on the other cases here. Thus, by the definition of Jeffrey conditionalization:

$$\begin{aligned} q(- \mid R \wedge H) &= \frac{q(- \wedge R \wedge H)}{q(R \wedge H)} \\ &= \frac{p(- \wedge R \wedge H \mid R \to H) \cdot q(R \to H) + p(- \wedge R \wedge H \mid R \to \neg H) \cdot q(R \to \neg H)}{p(\neg R \mid R \to H) \cdot q(R \to H) + p(\neg R \mid R \to \neg H) \cdot q(R \to \neg H)} \\ &= \frac{p(- \wedge R \wedge H \mid R \to H) + p(H, \neg R \mid R \to \neg H)}{p(\neg R \mid R \to H) + p(\neg R \mid R \to \neg H)} \\ &= \frac{p(- \wedge R \wedge H)}{p(R \wedge H)} \\ &= p(- \mid R \wedge H) \end{aligned}$$

Proofs of Theorems

We now turn to the more general results, stated in §4.5. I'll start with Theorem 3. First, however, let me just state two bits of terminology. In what follows, I use 'PC' as a shorthand for 'probabilistic centering', where this is defined as:

$$p(A \wedge (A \rightarrow C)) = p(A \wedge C). \qquad \text{(Probabilistic Centering)}$$

I then use 'AI' as a shorthand for 'antecedent independence', where this is the following condition:

$$p(A \rightarrow C) = p(A \rightarrow C \mid A).$$

Proof of Theorem 3. For the left-to-right direction, suppose that $p(A \rightarrow C) = 1$. Then, every admissible sequence of worlds is such that the first A-world is a C-world. This can be the case, however, only if there are no worlds in the context in which A is false C is true. But this implies that $A \subset C$ is true at each possible world, and so $p(A \supset C) = 1$.

Now, for the right-to-left direction, suppose that $p(A \supset C) = 1$. Then $p(A \wedge \neg C) = 0$. But by our assumption that p assigns positive credence to each world in the context, that means that every A-world in the context is a C-world. It then follows that, for every admissible sequence, the first A-world in that sequence is a C-world, which means that $p(A > C) = 1$. □

Proof of Theorem 4. Bradley (2005) shows that q comes from p by Adams conditionalization if and only if the following conditions are satisfied:

- **Antecedent Independence**. $q(A) = p(A)$
- **Rigidity**. The following conditions all hold:
  - $q(- \mid A, C) = p(- \mid A, C)$,
  - $q(- \mid A, \neg C) = p(- \mid A, \neg C)$,
  - $q(- \mid \neg A) = p(- \mid \neg A)$.

Thus, to prove the theorem, we only need to show that q satisfies these conditions.

Start with Independence. Since q comes from p by Jeffrey conditionalization on $\{A \rightarrow C, A \rightarrow \neg C\}$, we have:

$$\begin{aligned} q(A) &= q(A \rightarrow C) \cdot p(A \mid A \rightarrow C) + q(A \rightarrow \neg C) \cdot p(A \mid A \rightarrow \neg C) && \text{(J-Cond)} \\ &= p(A \rightarrow C) \cdot cr(A) + p(A \rightarrow \neg C) \cdot p(A) && \text{(AI)} \\ &= p(A). \end{aligned}$$

So Independence holds.

Now turn to Rigidity. In the first case:

$$
\begin{aligned}
q(- \mid A, C) &= \frac{q(-, A, C)}{q(A, C)} && \text{(Ratio)} \\
&= \frac{q(-, A, C, A \to C)}{q(A, C, A \to C)} && \text{(PC)} \\
&= \frac{q(-, A, C \mid A \to C) \cdot q(A \to C)}{q(A, C \mid A \to C) \cdot q(A \to C)} && \text{(Ratio)} \\
&= \frac{q(-, A, C \mid A \to C)}{q(A, C \mid A \to C)} && \text{(Algebra)} \\
&= \frac{p(-, A, C \mid A \to C)}{p(A, C \mid A \to C)} && \text{(J-Cond)} \\
&= \frac{p(-, A, C, A \to C)/p(A \to C)}{p(A, C, A \to C)/p(A \to C)} && \text{(Ratio)} \\
&= \frac{p(-, A, C, A \to C)}{p(A, C, A \to C)} && \text{(Algebra)} \\
&= \frac{p(-, A, C)}{p(A, C)} && \text{(PC)} \\
&= p(X \mid A, C) && \text{(Ratio)}
\end{aligned}
$$

The other two cases are proved similarly. So Rigidity holds as well. □

Proof of Corollary 1. Immediate, since, in this case, conditionalization is the special case of Jeffrey conditionalization in which $q(A \to C) = 1$, and the version of Adams conditionalization above is the special case of the more general version, in which $q(C \mid A) = 1$. □

Bibliography

Adams, E. W. (1966). Probability and the logic of conditionals. *Studies in Logic and the Foundations of Mathematics*, 265–316. https://doi.org/10.1016/s0049-237x(08)71673-2

Adams, E. W. (1970). Subjunctive and indicative conditionals. *Foundations of Language*, *6*(1), 89–94. http://www.jstor.org/stable/25000429

Adams, E. W. (1975). *The logic of conditionals*. Springer Netherlands. https://doi.org/10.1007/978-94-015-7622-2

Ahmed, A. (2013). Causal decision theory: A counterexample. *Philosophical Review*, *122*(2), 289–306. https://doi.org/10.1215/00318108-1963725

Ahmed, A. (2014a). Causal decision theory and the fixity of the past. *The British Journal for the Philosophy of Science*, *65*(4), 665–685. https://doi.org/10.1093/bjps/axt021

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

Ahmed, A. (2021). *Evidential decision theory*. Cambridge University Press.

Ahmed, A., & Spencer, J. (2020). Objective value is always newcombizable. *Mind*, *129*(516), 1157–1192. https://doi.org/10.1093/mind/fzz070

Albert, D. Z. (2000). *Time and chance*. Harvard University Press.

Bacon, A. (2015). Stalnaker's thesis in context. *The Review of Symbolic Logic*, *8*(1), 131–163. https://doi.org/10.1017/S1755020314000318

Bennett, J. (1984). Counterfactuals and temporal direction. *The Philosophical Review*, *93*(1), 57. https://doi.org/10.2307/2184413

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford University Press.

Bovens, L. (2009). Judy benjamin is a sleeping beauty. *Analysis*, *70*(1), 23â□□26. https://doi.org/10.1093/analys/anp127

Bovens, L., & Ferreira, J. L. (2010). Monty hall drives a wedge between judy benjamin and the sleeping beauty: A reply to bovens. *Analysis*, *70*(3), 473â□□481. https://doi.org/10.1093/analys/anq020

Boylan, D. (MS). *Evidence and conditional propositions* [Unpublished Manuscript].

Boylan, D., & Schultheis, G. (2021). How strong is a counterfactual? *Journal of Philosophy*, *118*(7), 373–404. https://doi.org/10.5840/jphil2021118728

Braddon-Mitchell, D. (2001). Lossy laws. *Noûss*, *35*(2), 260–277. https://doi.org/10.1111/0029-4624.00296

Bradley, R. (2000). A preservation condition for conditionals. *Analysis*, *60*(3), 219–222. https://doi.org/10.1111/1467-8284.00228

Bradley, R. (2005). Radical probabilism and bayesian conditioning∗. *Philosophy of Science*, *72*(2), 342–364. https://doi.org/10.1086/432427

Bradley, R. (2012). Multidimensional possible-world semantics for conditionals. *The Philosophical Review*, *121*(4), 539–571. https://doi.org/10.1215/00318108-1630921

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

Bradley, R., & List, C. (2009). Desire-as-belief revisited. *Analysis*, *69*(1), 31–37. https://doi.org/10.1093/analys/ann005

Bradley, R., & Stefánsson, H. O. (2017). Counterfactual desirability. *British Journal for the Philosophy of Science*, *68*(2), 485–533. https://doi.org/10.1093/bjps/axv023

Bradley, R., & Stefánsson, H. O. (2016). Desire, expectation, and invariance. *Mind*, *125*(499), 691–725. https://doi.org/10.1093/mind/fzv200

Broome, J. (1991). Discussion: Desire, belief, and expectation. *Mind*, *100*(398), 265–267. https://doi.org/10.1093/mind/c.398.265

Byrne, A., & Hájek, A. (1997). David hume, david lewis, and decision theory. *Mind*, *106*(423), 411–728. https://doi.org/10.1093/mind/106.423.411

Ciardelli, I., & Ommundsen, A. (2022). Probabilities of conditionals: Updating adams. *Noûs*. https://doi.org/10.1111/nous.12437

Collins, J. (1988). Belief, desire, and revision. *Mind*, *97*(387), 333–342.

Collins, J. (2015). Decision theory after lewis. In B. Loewer & J. Schaffer (Eds.), *A companion to david lewis* (pp. 446–458). John Wiley; Sons.

Costa, H. A., Collins, J., & Levi, I. (1995). Desire-as-belief implies opinionation or indifference. *Analysis*, *55*(1), 2–5. https://doi.org/10.1093/analys/55.1.2

Diaconis, P., & Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association*, *77*(380), 822–830.

Dorr, C. (2016). Against counterfactual miracles. *Philosophical Review*, *125*(2), 241–286. https://doi.org/10.1215/00318108-3453187

Dorr, C., & Hawthorne, J. (MS). *If…: A theory of conditionals* [Unpublished Manuscript].

Dorr, C., & Mandelkern, M. (MS). *The logic of sequences* [Unpublished Manuscript].

Douven, I. (2012). Learning conditional information. *Mind &amp Language*, *27*(3), 239–263. https://doi.org/10.1111/j.1468-0017.2012.01443.x

Douven, I., & Dietz, R. (2011). A puzzle about stalnaker?s hypothesis. *Topoi*, *30*(1), 31–37. https://doi.org/10.1007/s11245-010-9082-3

Douven, I., & Romeijn, J.-W. (2011). A new resolution of the judy benjamin problem. *Mind*, *120*(479), 637–670. https://doi.org/10.1093/mind/fzr051

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329. https://doi.org/10.1093/mind/104.414.235

Edgington, D. (2003). Counterfactuals and the benefit of hindsight. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world*. Routledge.

Edgington, D. (2021). Suppose and tell: The semantics and heuristics of conditionals: Timothy williamson. oxford: Oxford university press, 2020. viii + 278 pp. £30.00. isbn 978-0-19-886066-2. *History and Philosophy of Logic*, *43*(2), 188–195. https://doi.org/10.1080/01445340.2021.1958648

Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *68*(S3), S313–S324. https://doi.org/10.1086/392918

Elga, A. (2022). Confession of a causal decision theorist. *Analysis*. https://doi.org/10.1093/analys/anab040

Eva, B., Hartmann, S., & Rad, S. R. (2019). Learning from conditionals. *Mind*, *129*(514), 461–508. https://doi.org/10.1093/mind/fzz025

Fitelson, B. (2015). The strongest possible lewisian triviality result. *Thought: A Journal of Philosophy*, *4*(2), 69–74. https://doi.org/10.1002/tht3.159

Fusco, M. (2022). Dutch-booking indicative conditionals. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12910

Fusco, M. (forthcoming). Absolution of a causal decision theorist. *Noûs*.

Gallow, J. D. (2022). Causal counterfactuals without miracles or backtracking. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12925

Gibbard, A. (1981). Two recent theories of conditionals. In W. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 211–247). Reidel.

Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and applications of decision theory* (pp. 125–162). D. Reidel.

Gillies, A. S. (2007). Counterfactual scorekeeping. *Linguistics and Philosophy*, *30*(3), 329–360. https://doi.org/10.1007/s10988-007-9018-6

Goldstein, S., & Santorio, P. (2021). Probability for epistemic modalities. *Philosophers' Imprint*, *21*(33).

Goodman, J. (2014). Knowledge, counterfactuals, and determinism. *Philosophical Studies*, *172*(9), 2275–2278. https://doi.org/10.1007/s11098-014-0409-6

Gregory, A. (2017). Might desires be beliefs about normative reasons? In J. Deonna & F. Lauria (Eds.), *The nature of desire* (pp. 201–217). Oxford University Press.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge: Harvard University Press.

Groenendijk, J. A. G., & Stokhof, M. J. B. (1984). *Studies on the semantics of questions and the pragmatics of answers* (Doctoral dissertation). University of Amsterdam.

Günther, M. (2018). Learning conditional information by jeffrey imaging on stalnaker conditionals. *Journal of Philosophical Logic*, *47*(5), 851–876. https://doi.org/10.1007/s10992-017-9452-z

Hájek, A. (1989). Probabilities of conditionals—revisited. *Journal of Philosophical Logic*, *18*(4), 423–428. https://doi.org/10.1007/bf00262944

Hájek, A. (2012). The fall of "adams' thesis"? *Journal of Logic, Language and Information*, *21*(2), 145–161. https://doi.org/10.1007/s10849-012-9157-1

Hájek, A. (2015). On the plurality of lewis's triviality results. *A companion to David Lewis*, 425–445.

Hájek, A. (MS). *Most counterfactuals are false* [Unpublished Manuscript].

Hájek, A., & Hall, N. (1994). The hypothesis of the conditional construal of conditional probability. In E. Eells, B. Skyrms, & E. W. Adams (Eds.), *Probability and conditionals: Belief revision and rational decision* (p. 75). Cambridge University Press.

Hájek, A., & Pettit, P. (2004). Desire beyond belief. *Australasian Journal of Philosophy*, *82*(1), 77–92. https://doi.org/10.1080/713659805

Hamblin, C. L. (1973). Questions in Montague english. *Foundations of Language*, *10*(1), 41–53.

Hawthorne, J. (2005). Chance and counterfactuals. *Philosophy and Phenomenological Research*, *70*(2), 396–405. http://www.jstor.org/stable/40040799

Hedden, B. (2023). Counterfactual decision theory. *Mind*, *132*(527), 730–761. https://doi.org/10.1093/mind/fzac060

Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous*, *39*(4), 632–657. https://doi.org/10.1111/j.0029-4624.2005.00542.x

Higginbotham, J. (1986). Linguistic theory and davidson's program in semantics. In E. LePore (Ed.), *Truth and interpretation: Perspectives on the philosophy of donald davidson* (pp. 29–48). Cambridge: Blackwell.

Higginbotham, J. (2003). Conditionals and compositionality. *Philosophical Perspectives*, *17*(1), 181–194. https://doi.org/10.1111/j.1520-8583.2003.00008.x

Hoek, D. (2019). *The web of questions the web of questions: Inquisitive decision theory and the bounds of rationality* (Doctoral dissertation). New York University.

Hoek, D. (2022). Questions in action. *The Journal of Philosophy*, *119*(3), 113–143. https://doi.org/10.5840/jphil202211938

Holguín, B., & Teitel, T. (MS). *On the plurality of counterfactuals* [Unpublished Manuscript].

Ichikawa, J. (2011). Quantifiers, knowledge, and counterfactuals. *Philosophy and Phenomenological Research*, *82*(2), 287–313.

Ippolito, M. (2016). How similar is similar enough? *Semantics and Pragmatics*, *9*, 6–1.

Jackson, F. (1977). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, *55*(1), 3–21. https://doi.org/10.1080/00048407712341001

Jeffrey, R. C. (1965). *The logic of decision.* 1st Edition, University of Chicago Press.
Jeffrey, R. C. (1983). *The logic of decision.* 2nd Edition, Chicago; London: University of Chicago Press.
Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.
Joyce, J. M. (2009a). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 263–297). Synthese.
Joyce, J. M. (2009b). Causal reasoning and backtracking. *Philosophical Studies*, *147*(1), 139–154. https://doi.org/10.1007/s11098-009-9454-y
Joyce, J. M. (2016). Review of Arif Ahmed: Evidence, Decision and Causality. *Journal of Philosophy*, *113*(4), 224–232. https://doi.org/10.5840/jphil2016113413
Joyce, J. M. (2018). Deliberation and stability in Newcomb problems and pseudo-Newcomb problems. In A. Ahmed (Ed.), *Newcomb's problem* (pp. 138–159). Cambridge University Press. https://doi.org/10.1017/9781316847893.008
Joyce, J. (2004). Williamson on evidence and knowledge. *Philosophical Books*, *45*(4), 296–305. https://doi.org/10.1111/j.1468-0149.2004.0356c.x
Kaufmann, S. (2004). Conditioning against the grain. *Journal of Philosophical Logic*, *33*(6), 583–606. https://doi.org/10.1023/b:logi.0000046142.51136.bf
Kaufmann, S. (2005). Conditional predictions. *Linguistics and Philosophy*, *28*(2), 181–231. https://doi.org/10.1007/s10988-005-3731-9
Kaufmann, S. (2015). Conditionals, conditional probabilities, and conditionalization. In H.-C. Schmitz & H. Zeevat (Eds.), *Bayesian natural language semantics and pragmatics* (pp. 71–94). Springer.
Kelley, M. et al. (MS). *Accuracy and epistemic modalities* [Unpublished Manuscript].
Khoo, J. (2016). Probabilities of conditionals in context. *Linguistics and Philosophy*, *39*(1), 1–43. https://doi.org/10.1007/s10988-015-9182-z
Khoo, J. (2017). Backtracking counterfactuals revisited. *Mind*, fzw005. https://doi.org/10.1093/mind/fzw005
Khoo, J. (2022). *The meaning of 'If'*. New York, USA: Oxford University Press.
Khoo, J., & Mandelkern, M. (2018). Triviality results and the relationship between logical and natural languages. *Mind*, *128*(510), 485–526. https://doi.org/10.1093/mind/fzy006
Khoo, J., & Santorio, P. (2018). *Lecture notes: Probabilities of conditionals in modal semantics* [Unpublished Manuscript].
Kment, B. (2006). Counterfactuals and explanation. *Mind*, *115*(458), 261–310. https://doi.org/10.1093/mind/fzl261
Kment, B. (2014). *Modality and explanatory reasoning*. Oxford University Press.
Kment, B. (2023). Decision, causality, and predetermination. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12935
Lange, M. (2000). *Natural laws in scientific practice*. Oxford University Press.
Lewis, D. (1973a). Causation. *The Journal of Philosophy*, *70*(17), 556. https://doi.org/10.2307/2025310
Lewis, D. (1973b). *Counterfactuals*. Blackwell.
Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, *85*(3), 297. https://doi.org/10.2307/2184045
Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476. https://doi.org/10.2307/2215339
Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*(1), 5–30. https://doi.org/10.1080/00048408112340011
Lewis, D. (1986). *Philosophical papers: Volume II*. Oxford University Press.
Lewis, D. (1988a). Desire as belief. *Mind*, *97*(387), 323–332. https://doi.org/10.1093/mind/xcvii.387.323
Lewis, D. (1988b). Relevant implication. *Theoria*, *54*(3), 161–174. https://doi.org/10.1111/j.1755-2567.1988.tb00716.x

Lewis, D. (1988c). Statements partly about observation. *Philosophical Papers*, *17*(1), 1–31. https://doi.org/10.1080/05568648809506282
Lewis, D. (1996). Desire as belief II. *Mind*, *105*(418), 303–313. https://doi.org/10.1093/mind/105.418.303
Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, *97*(4), 182–197. https://doi.org/jphil200497437
Loewer, B. (2007). Counterfactuals and the second law. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited*. Oxford University Press.
Mandelkern, M. (2018). Talking about worlds. *Philosophical Perspectives*, *32*(1), 298–325. https://doi.org/10.1111/phpe.12112
Mandelkern, M. (forthcoming). *Bounds: The dynamics of interpretation* [Unpublished Manuscript]. Oxford University Press.
Maudlin, T. (2007). *The metaphysics within physics*. Oxford University Press.
McGee, V. (1989). Conditional probabilities and compounds of conditionals. *The Philosophical Review*, *98*(4), 485. https://doi.org/10.2307/2185116
McNamara, C. (MS-a). *Actual value and indeterminacy in decision theory* [Unpublished Manuscript], University of Michigan, Ann Arbor.
McNamara, C. (MS-b). *Desire-as-belief in context* [Unpublished Manuscript].
McNamara, C. (MS-c). *Most counterfactuals are indeterminate* [Unpublished Manuscript].
McNamara, C., & Zhang, S. (MS). *Why (not) conditionalize?* [Unpublished Manuscript].
Misak, C. (2020). *Frank ramsey: A sheer excess of powers*. Oxford University Press.
Moss, S. (2012). On the pragmatics of counterfactuals. *Noûs*, *46*(3), 561–586. https://doi.org/10.1111/j.1468-0068.2010.00798.x
Moss, S. (2013). Subjunctive credences and semantic humility. *Philosophy and Phenomenological Research*, *87*(2), 251–278. https://doi.org/10.1111/j.1933-1592.2011.00550.x
Moss, S. (2015). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics*, *8*(5), 1–81. https://doi.org/10.3765/sp
Moss, S. (2018). *Probabilistic knowledge*. Oxford University Press.
Moss, S. (MS). *A synthesis view of counterfactuals* [Unpublished Manuscript].
Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of carl g. hempel* (pp. 114–146). Reidel.
Nute, D. (1980). Conversational scorekeeping and conditionals. *Journal of Philosophical Logic*, *9*(2). https://doi.org/10.1007/bf00247746
Oddie, G. (1994). Harmony, purity, truth. *Mind*, *103*(412), 451–472. https://doi.org/10.1093/mind/103.412.451
Oddie, G. (2001). Hume, the bad paradox, and value realism. *Philo*, *4*(2), 109–122. https://doi.org/10.5840/philo20014210
Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press UK.
Popper, K., & Miller, D. (1983). A proof of the impossibility of inductive probability. *Nature*, *302*(5910), 687–688.
Price, H. (1989). Defending desire-as-belief. *Mind*, *98*(389), 119–127. https://doi.org/10.1093/mind/xcviii.389.119
Ramsey, F. P. (1929). General propositions and causality. In D. Mellor (Ed.), *F.p. ramsey: Philosophical papers*. Cambridge University Press.
Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*. https://doi.org/10.3765/sp.5.6
Rooij, R. (1999). Gibbard's problem the context dependence of conditional statements. *Proceedings of the Dutch-German Workshop on Monotonic Reasoning*.
Rosen, G. (2017). Metaphysical relations in metaethics. In T. McPherson & D. Plunkett (Eds.), *The routledge handbook of metaethics* (pp. 151–169). Routledge.

Rothschild, D. (2011). i

$$do indicative conditionals express propositions?$$

/i

. *Noûs*, *47*(1), 49–68. https://doi.org/10.1111/j.1468-0068.2010.00825.x
Rothschild, D. (2021). Living in a material world: A critical notice ofsuppose and tell: The semantics and heuristics of conditionalsby timothy williamson. *Mind*, *132*(525), 208–233. https://doi.org/10.1093/mind/fzab049
Sandgren, A., & Williamson, T. L. (2020). Determinism, counterfactuals, and decision. *Australasian Journal of Philosophy*, *99*(2), 286–302. https://doi.org/10.1080/00048402.2020.1764073
Santorio, P. (2019). Interventions in premise semantics. *Philosophers' Imprint*, *19*.
Santorio, P. (2022). Path semantics for indicative conditionals. *Mind*, *131*(521), 59–98. https://doi.org/10.1093/mind/fzaa101
Schultheis, G. (2023). Counterfactual probability. *The Journal of Philosophy*, *120*(11), 581–614. https://doi.org/10.5840/jphil20231201133
Schultheis, G. (forthcoming). Counterfactual probability. *Journal of Philosophy*.
Skyrms, B. (1980a). *Causal necessity: A pragmatic investigation of the necessity of laws*. Yale University Press.
Skyrms, B. (1980b). The prior propensity account of subjunctive conditionals. *Ifs* (pp. 259–265). Springer.
Skyrms, B. (1982). Causal decision theory. *The Journal of Philosophy*, *79*(11), 695. https://doi.org/10.2307/2026547
Skyrms, B. (1984). *Pragmatics and empiricism*. Yale University Press, New Haven.
Slote, M. A. (1978). Time in counterfactuals. *Philosophical Review*, *87*(1), 3–27. https://doi.org/10.2307/2184345
Sobel, J. H. (1994). *Taking chances: Essays on rational choice*. Cambridge University Press.
Solomon, T. C. P. (2021). Causal decision theory's predetermination problem. *Synthese*, *198*(6), 5623–5654. https://doi.org/10.1007/s11229-019-02425-0
Solomon, T. C. P. (MS). *Libertarian decision theory* [Unpublished Manuscript].
Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory (american philosophical quarterly monographs 2)* (pp. 98–112). Oxford: Blackwell.
Stalnaker, R. (1970). Probability and conditionals. *Philosophy of Science*, *37*(1), 64–80. https://doi.org/10.1086/288280
Stalnaker, R. (1974). Pragmatic presuppositions. In R. Stalnaker (Ed.), *Context and content* (pp. 47–62). Oxford University Press.
Stalnaker, R. (1975). Indicative conditionals. *Philosophia*, *5*(3), 269–286. https://doi.org/10.1007/bf02379021
Stalnaker, R. (1978). Assertion. *Syntax and Semantics (New York Academic Press)*, *9*, 315–332.
Stalnaker, R. (1981a). A defense of conditional excluded middle. In W. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 87–104). Reidel.
Stalnaker, R. (1981b). Letter to David Lewis. In W. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 151–152). Reidel.
Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.
Stalnaker, R. (2021). Counterfactuals and probability. In L. Walters & J. Hawthorne (Eds.), *Conditionals, paradox, and probability: Themes from the philosophy of dorothy edgington*. Oxford University press.
Stalnaker, R. (MS). *Counterfactuals, compatibilism, and rational choice* [Unpublished Manuscript].
Stalnaker, R., & Jeffrey, R. (1994). Conditionals as random variables. In E. Eells, B. Skyrms, & E. W. Adams (Eds.), *Probability and conditionals: Belief revision and rational decision* (p. 31). Cambridge University Press.

Stalnaker, R., & Thomason, R. (1970). A semantic analysis of conditional logic. *Theoria*, *36*(1), 23–42. https://doi.org/10.1111/j.1755-2567.1970.tb00408.x

Steele, K., & Sandgren, A. (2020). Levelling counterfactual scepticism. *Synthese*, *199*(1-2), 927–947. https://doi.org/10.1007/s11229-020-02742-9

Stefánsson, H. O. (2014). Desires, beliefs and conditional desirability. *Synthese*, *191*(16), 4019–4035. https://doi.org/10.1007/s11229-014-0512-4

van Fraassen, B. (1976). Probabilities of conditionals. In W. H. C. Hooker (Ed.), *Foundations of probability theory, statistical inference, and statistical theories of science*.

van Fraassen, B. (1981). A problem for relative information minimizers in probability kinematics. *British Journal for the Philosophy of Science*, *32*(4), 375–379. https://doi.org/10.1093/bjps/32.4.375

van Fraassen, B., Hughes, R. I. G., & Harman, G. (1986). A problem for relative information minimizers, continued. *The British Journal for the Philosophy of Science*, *37*(4), 453–463. https://doi.org/10.1093/bjps/37.4.453

Vasudevan, A. (2020). Entropy and insufficient reason: A note on the judy benjamin problem. *The British Journal for the Philosophy of Science*, *71*(3), 1113–1141. https://doi.org/10.1093/bjps/axy013

Von Fintel, K. (2001). Counterfactuals in a dynamic context. *Current Studies in Linguistics Series*, *36*, 123–152.

Weintraub, R. (2007). Desire as belief, lewis notwithstanding. *Analysis*, *67*(2), 116–122. https://doi.org/10.1093/analys/67.2.116

Williams, R. (2012). Counterfactual triviality: A lewis-impossibility argument for counterfactuals. *Philosophy and Phenomenological Research*, *85*(3), 648–670. https://doi.org/10.1111/j.1933-1592.2012.00636.x

Williamson, T. (2020). *Suppose and tell: The semantics and heuristics of conditionals*. Oxford, England: Oxford University Press.

Williamson, T. L., & Sandgren, A. (forthcoming). Law-abiding causal decision theory. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/715103

Yalcin, S. (2016). Belief as question-sensitive. *Philosophy and Phenomenological Research*, *97*(1), 23–47. https://doi.org/10.1111/phpr.12330