Knowledge

A Human Interest Story

Brian Weatherson

June 14, 2019

Contents

Chapter 1

Prologue

What we know is sensitive to ever so many factors. The thesis of this book is that one of those factors is which questions we are interested in.

To know something in the context of answering a question requires, among other things, that we can treat the thing known as a fixed starting point for that inquiry. But there is practically nothing that we can treat as a fixed point in any inquiry whatsoever. If the stakes are high enough, or cost of questioning low enough, it can be appropriate to question almost anything at all. If what we know is what can be treated as a fixed starting point in any inquiry whatsoever, then we know almost nothing. Since we do know a lot, this can't be right.

The solution is to say that what we know varies depending on what questions we are interested in, and so on what inquiries we are engaged in. If a proposition can't be treated as a fixed point for a particular inquiry, then it can't be known so long as we are engaged in that inquiry. But it can, at least possibly, be known at other times.

The story of investigations into knowledge over the last fifty years is the story of finding ever more things that knowledge is sensitive to. The thesis of this book is that human interests, in particular the interests of the would be knower, should be added to that list.

It helps to begin with a list of some of the things which we already know that knowledge is sensitive to. Much of this list, indeed much of the rest of this chapter, will be familiar to experts. But for most readers, it helps to locate the view of the book by placing it in the context of familiar theses about what knowledge is sensitive to. So I'm going to describe a mundane case of knowledge, then discuss various ways in which that knowledge could be lost if the world were different.

Our protagonist, Charlotte, is reading a book about the build up to World War One. In the base case, the book is Christopher Clark's *The Sleepwalkers* (Clark, 2012), though in some of the variants we'll discuss she reads a less impressive book. In it she reads the remarkable story of Henriette Calliaux, the second wife of anti-war French politician Joseph Calliaux. As you may already know, Henriette Calliaux shot and killed Gaston Calmette, the editor of *Le Figaro*, after *Le Figaro* published a string of damaging articles about Joseph Calliaux. The murder took place on March 16, 1914, and the trial was that July. It ended on July 28 with her acquittal.

Charlotte reads all of this and believes it. And indeed it is true. And the book is reliable. Although Charlotte does believe what the book says about Henriette Calliaux, she is not credulous. She is an attentive enough, and skilled enough, reader of contemporary history to know when historians are likely to be going out on a limb, and when they are not being as clear as one might like in reflecting how equivocal the evidence is. But Clark is a good historian, and Charlotte is a good reader, and the beliefs she takes from the book are both true and supported by the underlying evidence.

Focus for now on this proposition

> Henriette Caillaux's trial for the murder of Gaston Calmette ended in her acquittal in late July 1914.

Call this proposition p. In this base case, Charlotte knows that p. But there are ever so many ways in which Charlotte could fail to have known it. The following three are particularly important .

> *Variant J*
>
> Charlotte didn't finish the book. She only got as far as the start of Caillaux's trial, but lost interest in the machinations of the diplomats in the late stages of the July crisis. Still, she had a strong hunch that Caillaux would be acquitted, and on this basis, firmly believed that she would be.

> *Variant T*
>
> Charlotte is in a world where things went just as in the actual world up to the trial, but then Caillaux was found guilty. Despite this, Charlotte reads a book that is word-for-word identical to Clark's book. That is, it falsely says that Caillaux

> was acquitted, before quickly moving back to talking about the war. Charlotte believes, falsely, that *p*.
>
> *Variant B*
>
> Charlotte reads the book to the end, but she can't believe that Caillaux would be acquitted. The evidence was conclusive, she thought. She is torn because she also can't really believe a historian would get such a clear fact wrong. But she also can't believe anyone would be acquitted in such a trial. So she withholds judgment on the matter, not sure what actually happened in Caillaux's trial.

In all three of these variants, Charlotte does not know that *p*. I take these three kinds of cases to be good evidence that knowledge requires (respectively) justification, truth, and belief. In variant J, Charlotte lacks knowledge because her belief in *p* is not justified; it is a mere hunch. In variant T, Charlotte lacks knowledge because her belief is not true; it is an honest mistake. In variant B, Charlotte lacks knowledge because she doesn't even believe *p*; she has the evidence, but does not accept it.

In all three cases there are philosophers who argue that these conditions are not strictly necessary, but it would take us too far afield to debate these points. I will simply take for granted that cases like Variant J show justification (or some kind of well-groundedness) is necessary for knowledge, Variant T shows that truth is necessary for knowledge, and Variant B shows belief (or at least some kind of strong acceptance) is necessary for knowledge. (I will come back to some of the issues about belief in chapter 3.)

For a short while in the mid-20th century, some philosophers thought these conditions were not merely necessary for knowledge, but jointly sufficient. To know that *p* just is to have a justified, true belief that *p*. This became known, largely in retrospect, as the JTB theory of knowledge. It fell out of fashion dramatically after a short but decisive criticism was published by Edmund Gettier (1963). But Gettier's criticism was not original; he had independently rediscovered a point made by the 8th century philosopher Dharmottara (Nagel, 2014). Here is a version of the kind of case Dharmottara discovered.

> *Variant D*
>
> Like in Variant J, Charlotte stops reading before the denouement. And she believes that Caillaux was acquitted. But she

> does not believe this on the basis of a hunch. Rather, she believes it because she read in another book that official France was so discombobulated in July 1914 that it didn't manage to convict a single murderer. This is false, but Charlotte used it to reason to the true conclusion that *p*.

In Variant D Charlotte does not know that *p*. She does not know it because reasoning that relies on a falsehood in just this way does not ground knowledge. So here is another thing that knowledge is sensitive to - whether the grounds for one's belief are true.

The variations from now on will not be as intuitively clear; whether Charlotte knows that *p* will be matters of greater dispute than in these first four variants. But I think there is a good case to be made in each of them. The first case is a version of an example due to Gilbert Harman (1973, 143ff).

> *Variant H*
>
> It is surprising that Charlotte has never heard of Henriette Caillaux, because in the world she inhabits, the story of Calliaux is infamous. She is as well known as other famous killers like Ned Kelly, Jack the Ripper, and Lee Harvey Oswald. Novels, plays and movies are frequently made about her killing of Calmette. But in all of these popular depictions, the ending is fictionalised. Every one of them ends with Caillaux's conviction and execution. This happened because the authorities were so embarrassed by her acquittal that they created a vast alternative reality in which Caillaux was convicted. Charlotte, by an amazing coincidence, is the only person to have not encountered this story. So when she reads a word-for-word duplicate of Clark's book, she doesn't realise it is controversial, and believes that *p*. Had she seen any of these books or plays, she would have assumed her book was making some mistake, since it is 'common knowledge' that Caillaux was convicted.

Intuitions may vary on this, but I don't think Charlotte knows that *p* in Variant H. If that's right, then whether Charlotte knows that *p* is sensitive not just to the evidence she has, but to the evidence that is all around her. If she's swimming in a sea of evidence against *p*, and by the sheerest luck has not run into it, that can block knowledge that *p*.

The previous example relied on the possibility of counter-evidence being everywhere. Possibly all that matters is that the counter-evidence is in just the right *somewhere*.

> *Variant S*
>
> In this world, an over-zealous copy-editor makes a last minute change to the very first printing of Clark's text. Not able to believe that Caillaux was acquitted - the evidence was so conclusive - they change the word 'acquittal' to 'conviction' in the sentence describing the end of the trial. Happily, this error is quickly caught, and it is only the very first printing of the book that contains the mistake. Charlotte started reading the book after seeing it in a second-hand shop. The shop had two copies: one from the flawed first printing, and one from a later printing. Charlotte buys the later one because it is the first one she sees; had she entered the history section from the other direction, she would have bought the first printing, and come to believe that *p* is false.

In this case, Charlotte doesn't know that *p* is true. There is too much luck in her happening to buy the later printing rather than the earlier printing. The belief-forming method that she uses - buy an apparently authoritative history book and believe the plausible and well-supported things it says - goes wrong on just this question in a very nearby possible world. (That is, it goes wrong in the world where she picks up the other copy.) And that kind of luck is incompatible with knowledge.

The contemporary terminology for this is that a belief forming method only yields knowledge if it is *safe*. And a method is safe only if it doesn't go wrong in nearby, realistic, scenarios (Williamson, 2000). So whether one knows is sensitive to not just the evidence one has, but the evidence one could easily have had.

But safety in this sense is a tricky notion. In Variant K, Charlotte seemingly does know that *p*.

> *Variant K*
>
> Charlotte detests reading books on paper, and only ever reads on her Kindle (an electronic book-reading device). Just like in Variant S, there was an error in the first printing of Clark's book. But the Kindle version never contained this error, and

> in any case, Kindle versions are updated frequently so even if it had, the error would have been quickly corrected. Charlotte reads the book on her Kindle, and comes to believe that *p*.

In Variant K, Charlotte does know that *p*. She believes *p* on good evidence from a trustworthy source, and there is no realistic possibility where she goes wrong on this question by trusting this source. I'll return to the difference between Variants S and K in a little, but first I want to consider two more cases.

> *Variant C*
>
> Charlotte reads Clark's book and believes *p*. But like in Variant B, she was sure that Caillaux would be convicted. And she still thinks it is absurd that someone would be acquitted given this evidence. But rather than responding to these conflicting pressures by withholding judgment, she responds by both believing that *p* is true, and believing it is false. She is just inconsistent, like so many of us are in so many ways.

It seems to me that in this case, Charlotte does not know that *p*. The incoherence in her beliefs on this very point undermines her claim to knowledge. And at last we get to the case that motivates this book.

> *Variant I*
>
> Charlotte reads the book, and believes that *p*. She is then offered a bet by a curiously benevolent deity. If she takes the bet, and *p* is true, she wins a dinner at her favourite bistro, Le Temps des cerises. If she takes the bet, and *p* is false, she is cast into The Bad Place for eternity. If she declines the bet, life goes on as normal. And now she's deciding what to do.

The main argument I'll defend in this book goes as follows.

1. Charlotte shouldn't take the bet because it's too risky.
2. If Charlotte knows that *p*, then she knows that taking the bet will have good consequences.
3. If Charlotte knows that taking the bet will have good consequences, then taking the bet is not too risky.
4. So, Charlotte doesn't know that *p*.

So being offered the bet, having her practical situation changed in this way, changes what Charlotte knows. When she read the book, she knew that *p*; now she doesn't. What she knows is sensitive to whether she is deliberating about questions like *Should I take this bet?*

The explanation for why her knowledge is interest-relative is that what she knows depends on what she can, rationally, take as a fixed starting point in inquiry. When inquiring about whether to take the bet, she can take as fixed that the terms of the bet include a nice dinner if *p* is true and disaster if it is false, that Le Temps des cerises exists, that The Bad Place is bad, and so on. To be sure, she shouldn't be perfectly certain of any of these things, but if we had to account for all of our uncertainty in every one of our deliberations, we wouldn't be able to do anything. So she takes those things as fixed. But she cannot take as fixed that *p* is true. The rational attitude to take towards *p*, in the context of this very inquiry, is that it is very very likely. And that's a different attitude to taking it as fixed and true. And that's why she can't know *p*, while doing this inquiry.

But there are other inquiries where she can, rationally, take quite a different attitude towards *p*. If someone asks her what was happening in France in July 1914, and she is weighing up how to answer, she can take it as given that *p* is true. That doesn't settle whether *p* should be part of her answer, since it might be insufficiently relevant. There were a lot of things happening in France in July 1914, like in every month in every country, and she has to decide what to leave in and what to leave out. But, at least in the normal case, she should make that decision while taking it as given that *p* is true.

So in the context of some inquiries, Charlotte knows that *p*, and in the context of other inquiries, she does not. What she knows varies with her inquiries. That is, it varies with what questions she is interested in. Her knowledge is interest-relative; it is sensitive to her interests.

But the point of this little survey is that this is just one more thing that knowledge is sensitive to. Variants J, T, B, D, H, S and C all showed other ways in which Charlotte's knowledge was sensitive to details of her situation. The aim of this book is that we shouldn't stop with variants like those seven; knowledge is sensitive to more than they show.

This does not mean that the analysis of knowledge should include references to interests, or indeed to anything else that covers the other seven variants where Charlotte loses knowledge. Philosophers used to think that just like we can chemically analyse jadeite as $NaAlSi_2O_6$, we can break it down into sodium, aluminium, silicon and oxygen, we can analytically

break knowledge down into its constituent parts. This is a much less popular view nowadays, largely due to the failure of attempts at analysis. But we don't have to think that there is any such analysis on offer to think that, for example, it is a necessary truth that whatever is known is true. Similarly, we don't have to think that there is any analysis of knowledge that makes reference to the interests of the knower in order to think that what one knows is sensitive to what one's interests are.

In listing a number of different cases in which Charlotte can lose knowledge of *p*, I do not mean to suggest that the explanation of why Charlotte loses knowledge is different in each case. there is no common explanation of multiple cases. Indeed, in many of these cases her loss of knowledge is explained in the same way. Linda Zagzebski (1994) argued, convincingly, that any theory of knowledge that gave distinct a explanation of cases like Variant T (where the belief is well-grounded, but false) from its explanation of the other cases would end up with insuperable difficulties.So there must be some overlap between the explanations of the cases.

And indeed, I won't treat Variant I as independent of the rest. There are close connections between what's going on in Variant C and what's going on in Variant I. Assume Charlotte holds on to her belief in *p*, and treats it as a fixed point in her reasoning. Then consider her attitudes to these three propositions.

1. *p*
2. If *p*, taking the bet has only good consequences.
3. Taking the bet is irrational, because it might have awfully bad consequences.

If she accepts all three of those claims, she is incoherent. And this incoherence, like in Variant C, undermines her claim to knowledge that *p*. If she rejects 1, then clearly she doesn't know *p*. If she rejects 2, she hasn't understood the bet she is facing. (And that might mean she keeps knowledge that *p*; I'll return to that kind of case later in the book.) If she rejects 3, then she's irrational. And if the only way to save the coherence of belief in *p* is through being irrational elsewhere, that also defeats a claim to knowledge.

So I think the interest-relativity of belief is best thought of as a side-effect of the way in which knowledge is required to be coherent with the rest of one's mental states. It isn't some free-standing addition to the theory of knowledge.

Coherence is a tricky notion though. We don't want to say that anyone who is incoherent about anything doesn't know anything. All of us have some incoherencies, so this would amount to wholesale scepticism. We have to be careful in saying which incoherencies undermine which claims to knowledge. It is just at this point that ultimately the interest-relativity really comes in. Whether a particular incoherence undermines knowledge that p is a function of whether the would-be knower is interested in the subject matter of the incoherent states.

There is something special about the first three variants: Variants J, T and B. The three conditions on knowledge they reveal have a special explanatory role. If someone knows something, that is explained (at least in large part) by the thing's being true, and being believed, and this belief being justified. Whether they know the thing is sensitive to all sorts of other different factors, including their interests. But arguably none of those other factors explain why one has knowledge in the cases that one does.

The distinction I'm relying on here, that something can be counterfactually relevant to whether a state obtains without being part of the explanation of that state, is easier to understand in more mundane cases of explanation. It is, famously, hard to explain the origins of World War One. But without settling all the causal and explanatory issues about the war's origins, we can confidently make the following two claims.

C Had a giant asteroid struck Sarajevo on June 27, 1914, the war would not have started when it did.

NE It is no part of the explanation of the start of the war that no such giant asteriod struck Sarajevo on June 27, 1914.

The counterfactual claim C can easily be verified by thinking about the consequences of giant asteroid strikes. (See, for example, the extinction of the dinosaurs.) And the claim about explanation NE can be verified by thinking about how absurd the task of explanation would be if it were false. For every possible event that could have changed history, but didn't, we'd have to include its non-happening in our explanation of the war. The non-occurrence of every possible alien invasion, mass pandemic, or tulip mania that could have happened, and would have made a difference, would be part of our explanation. This seems absurd too.

So the origins of the war are sensitive to whether there was a giant asteroid strike, but the lack of a giant asteroid strike is no part of the explanation for why the war took place. I want to say the same thing

about knowledge and interests. What one knows is always (in principle) sensitive to what one's interests are. But in cases where one knows, one's knowledge is rarely explained by what one's interests are.

It is often thought to be wildly implausible that interests should matter to knowledge in the way that I'm arguing. I think that some of what people are reacting to here is the implausibilty of thinking interests could explain knowledge. And that is, indeed, implausible. But sensitivity doesn't imply explanatoriness, and I'm only defending the sensitivity claim.

Most of the people who think that it is implausible that interests matter to knowledge are happy acknowledging the varieties of sensitivity that are revealed by Variants J, T, B, D, H, S and C. They just think this one new kind of sensitivity is a bridge too far. There are a number of challenges to this kind of positiion, but I'll end this introduction by discussing a little one of the most mundane. What adjective do you use to cover the factors that those seven variants (and others like them) suggest knowledge is sensitive to, while excluding factors like interests

One option is to call them the 'traditional' factors. Now since discussion of, say, safety only really became widespread in the 1990s, the tradition of including it in one's theory of knoweldge is quite a new one. But I don't mind calling new things traditional. I'm Australian, and we have great traditions like the Essendon-Collingwood traditional Anzac Day match, which also dates to the 1990s. But this terminology has a very short shelf-life. After all, we've been discussing the role of interests in epistemology since at least 2002 (Fantl and McGrath, 2002), so that's almost long enough to be traditional as well.

Another option is to say that they are the factors that are truth-connected, or truth-relevant. But there's no way to make sense of this notion in a way that gets at what is wanted. For one thing, it's really not obvious that coherence constraints (like we need for Variant C) are connected to truth. For another, all Variant I suggests is that we need a principle like the following in our theory of knowledge.

> Someone knows something only if their evidence is strong enough for them to rationally treat the thing as a fixed starting point in their inquiries.

On the face of it, that's truth-connected. It says knowledge requires strong evidence. Now, of course, it also says just how strong a requirement is depends on what their inquiries are. But what the critics want to say is

not just that every factor that matters to knowledge is truth-relevant, but every aspect of every factor that matters is truth-relevant.

And this last claim is independently implausible. Think about the difference between Variant S and Variant K. (Similar cases are used, though to make slightly different points, in Gendler and Hawthorne (2005).) If you think Charlotte knows that *p* in Variant K, but not in Variant S, then you think that whether she knows that *p* might depend on whether she prefers reading books on paper or on an electronic device. And that preference isn't truth-relevant or truth-connected. Or compare Variant H to a case that is just like it, except a while ago Charlotte emigrated to a country where no one ever talks about Henriette Caillaux. In the variant, Charlotte knows that *p*. So her knowledge of French history is sensitive to her emigration status. And that isn't truth-relevant or truth-connected.

If we think knowledge is sensitive in any way to the features of one's environment, then it will end up being sensitive to one's interests. (This point is not new; it is well made by Nilanjan Das.) The notion of 'environment' that matters to these constraints can't be given a narrowly physical definition. It can't be defined, for instance, as all places within some specific distance from me. My environment, in the relevant sense, consists of a network of towns and universities throughout the globe, and excludes any number of places a short drive away. But should I become more interested in nearby suburbs than far away colleges, my environment would change. Since what is in one's environment is interest-relative, then any theory that accounts for cases like Variants S or H by appeal to features of Charlotte's environment is an interest-relative theory. And most 'traditional', 'truth-connected' theories of knowledge do account for cases like Variants S and H in this way.

Now this isn't the only way, or even the main way, that interests matter to knowledge. But it is useful to see how easy it is for factors like interests to become relevant to knowledge. And when we return later in the book to objections to my version of the interest-relative theory, it will be useful to bear in mind just how many weird and wonderful things knowledge is sensitive to.

Chapter 2

Interests in Epistemology

2.1 Red or Blue?

The key argument that knowledge is interest-relative starts with a puzzle about a game. Here are the rules of the game, which I'll call the Red-Blue game.

1. Two sentences will be written on the board, one in red, one in blue.
2. The player will make two choices.
3. First, they will pick a colour, red or blue.
4. Second, they say whether the sentence in that colour is true or false.
5. If they are right, they win. If not, they lose.
6. If they win, they get $50, and if they lose, they get nothing.

Our player is Anisa. She has been reading some medieval history, and last night was reading about the Battle of Agincourt. She was amused to see that it too place on her birthday, October 25, and in 1415, precisely 595 years before her own birthday. The book says all these things about the Battle of Agincourt because they are actually true, and when she read the book, Anisa believed them. She believed them because she had lots of independent evidence that the book was reliable (it came from a respected author and publisher, it didn't contradict her well-grounded background beliefs), and she was sensitive to that evidence of its reliability. And, indeed, the book was generally reliable, as well as accurate on this point.

Anisa comes to know that she is playing the Red-Blue game, and that these are its rules. She does not come to know any other relevant fact about the game. When the game starts, the following two sentences are written on the board, the first in red, the second in blue.

- Two plus two equals four.
- The Battle of Agincourt took place in 1415.

Anisa looks at this, thinks to herself, "Oh, my book said that the Battle of Agincourt was in 1415, so (given the rules of the game) playing Blue-True will be as good as any other play, so I'm playing Blue-True. Playing Red-True would get the same amount, since obviously two plus two is four, but I'm going to play Blue-True instead". And that's what she does, and she wins the $50.

Intuitively, Anisa's move here is irrational. It doesn't cost her anything - she gets the $50. And it's not that irrational as these things go - she costs herself $50 in the somewhat distant worlds where her reliable book gets this fact wrong. But it was still irrational. She took a needless risk, when there was a simple safe option on the table.

I'm going to argue, at some length, that the best explanation of why it is irrational for Anisa to play Blue-True is that knowledge is interest-relative. Given her interests in learning about late medieval history, when she was at home reading the book, Anisa knew that the Battle of Agincourt did take place in 1415. Given her interests in playing this game well, Anisa does not know this. When she is moved into the game situation, she loses some knowledge she previously had.

Interest-relativity is often taken to be a wild and radical development in the theory of knowledge. And it is certainly a reform. It's not a new reform proposal; both the proposal, and many of the details, are set out in works by Jeremy Fantl and Matthew McGrath (2002; 2009), John Hawthorne (2004), and Jason Stanley (2005). But relative to the epistemological status quo circa 1990, it is different. But then again, factors that were widely held to affect knowledge according to the status quo of either today or of 1990 would have seemed wild and radical relative to the epistemological status quo circa 1960. The factors that make a belief safe, or sensitive, or reliable, or undefeated, were well outside the realms of factors that late 20th century epistemologists thought relevent to knoweldge. There are many things that are irrelevant to how probable a belief is that are relevant to whether it is knowledge, as the epistemological literature of the late 20th Century makes clear. The proposal here is that interests are one more addition to this motley bunch.

The standard arguments for and against interest-relativity to date have not focussed on examples like Anisa's, but on examples like Blaise that I'll present shortly. There are exceptions. The structure of Anisa's example is similar, in the features that matter to me, to the examples of low-cost

checking that Bradley Armour-Garb (2011) discusses. (Though he draws contextualist conclusions from these examples, not interest-relative ones.) And it is similar to some of the cases of three-way choice that Charity Anderson and John Hawthorne deploy in arguing against interest-relativity (2019a; 2019b). But mostly people have focussed on cases like the following.

Last night, Blaise was reading the same book that Anisa was reading. And he too was struck by the fact that the Battle of Agincourt took place on October 25, 1415. Today he is visited by a representative of the supernatural world, and offered the following bet. (Blaise knows these are the terms of the bet, and doesn't know anything else relevant.) If he declines the bet, life will go on as normal. If he accepts, one of two things will happen.

- If it is true that the Battle of Agincourt took place in 1415, an infant somewhere will receive one second's worth of pure joy, of the kind infants often get playing peek-a-boo.
- If it is false that the Battle of Agincourt took place in 1415, all of humanity will be cast into The Bad Place for all of eternity.

Blaise takes the bet. The Battle of Agincourt was in 1415, and he can't bear the thought of a lovable baby missing that second of pure joy.

Again, there is an intuition that Blaise did something horribly wrong here. And this intuition is best explained, I will argue, by letting knowledge be interest-relative. But the argument that the interest-relativity of knowledge is the best explanation of what's going on is, in my view, somewhat weaker in Blaise's case than in Anisa's. It's not that I think the interest-relative explanation of the case is wrong; in fact I think it's basically correct. It's rather that there are somewhat more plausible interest-invariant explanations of Blaise's case than of Anisa's. So I'll focus on Anisa, not Blaise.

This choice of focus occasionally means that this book is less connected to the existing literature than I would like. I occasionally infer what a philosopher would say about cases like Anisa from what they have said about cases like Blaise. And I suspect in some cases I'll get those inferences wrong. But I want to set out the best argument for the interest-relativity of knowledge that I know, and that means going via the example of Anisa.

Though I am starting with an example, and with an intuition about it, I am not starting with an intuition about what is known in the example. I don't have any clear intuitions about what Anisa knows or doesn't know

while playing the Red-Blue game. The intuition that matters here is that her choice of Blue-True is irrational. It's going to be a matter of inference, not intuition, that Anisa lacks knowledge.

And that inference will largely be by process of elimination. In section 2.2 I will set out four possible things we can say about Anisa, and argue that one of them must be true. (The argument won't appeal to any principles more controversial than the Law of Excluded Middle.) But all four of them, including the interest-relative view I favour, have fairly counterintuitive consequences. So something counterintuitive is true around here. And this puts a limit on how we can argue. At least one instance of the argument *this is counterintuitive, so it is false* must fail. And that casts doubt over all such arguments. This is a point that critics of interest-relativity haven't sufficiently acknowledged, but it also puts constraints on how one can defend interest-relativity.

When Anisa starts playing the Red-Blue game, her practical situation changes. So you might think I've gone wrong in stressing Anisa's interests, not her practical situation. I've put the focus on interests for two reasons. One is that if Anisa is totally indifferent to money, then there is no rational requirement to play Red-True. We need to posit something about Anisa's interests to even get the data point that the interest-relative theory explains. The second reason, which I'll talk about more in section 2.5, is that sometimes we can lose knowledge due to a change not in our practical situation, but our theoretical interests.

In the existing literature, views like mine are sometimes called versions of **subject-sensitive invariantism**, since they make knowledge relevant to the stakes and salient alternatives available to the subject. But this is a bad name; of course whether a knowledge ascription is true is sensitive to who the subject of the ascription is. I know what I had for breakfast and you (probably) don't. What is distinctive is which features of the subject's situation that the interest-relative theory says are relevant, and calling it the interest-relative theory of knowledge makes it clear that it is the subject's interests. In the past, I've called it **interest-relative invariantism**. But, for reasons I'll say more about in section 2.7, I'm not committed to *invariantism* in this book. So it's just the interest-relative theory, or IRT.

2.2 Four Families

A lot of philosophers have written about cases like Anisa and Blaise over the last couple of decades. Relatedly, there are a huge number of theories

that have been defended concerning these cases. Rather than describe them all, I'm going to start with a taxonomy of them. The taxonomy has some tricky edge cases, and it isn't always trivial to classify a philosopher from their statements about the cases. But I find it a helpful way to start thinking about the possible moves available.

Our first group of theories are the **sceptical** theories. They deny that Anisa ever knew that the Battle of Agincourt was in 1415. The particular kind of sceptic I have in mind says that if someone's epistemic position is, all things considered, better with respect to q than with respect to p, that person doesn't know that p. The core idea for this sceptic, which perhaps they draw from work by Peter Unger (1975), is that knowledge is a maximal epistemic state, so any non-maximal state is not knowledge. The sceptics say that for almost any belief, Anisa's belief that two plus two is four will have higher epistemic standing than that belief, so that belief doesn't amount to knowledge.

Our second group of theories are what I'll call **epistemicist** theories. The epistemicists say that Anisa's reasoning, and perhaps Blaise's reasoning too, is perfectly sound. They both know when the Battle of Agincourt took place, so they both know that the choices they take are optimal, so they are rational in taking those choices. The intuitions to the contrary are, say the epistemicist, at best confused. There is something off about Anisa and Blaise, perhaps, but it isn't that these particular decisions are irrational.

It's not essential to epistemicism as I'm construing it, but I think by far the most plausible form of epistemicism takes on board Maria Lasonen-Aarnio's point that act-level and agent-level assessments might come apart.[1] Taking the bet reveals something bad about Blaise's character, and arguably manifests a vice, but the act itself is rational. It's that last claim, that the actions are rational, that is distinctive of epistemicism as I'm understanding it.

The third group is the **pragmatist** theories, and this group includes the interest-relative theory that I'll defend. The pragmatists say that Anisa did know when the Battle of Agincourt was, but now she doesn't. The change in her practical situation, combined with her interest in getting

[1] See Lasonen-Aarnio (2010; 2014) for more details on her view. In *Normative Externalism*, I describe the difference between act-level and agent-level assessments as the difference between asking whether what Anisa does is rational, and whether Anisa's action manifests wisdom (Weatherson, 2019, 124-5). The best form of epistemicism, I'm suggesting, says that Anisa and Blaise are rational but unwise. This isn't Lasonen-Aarnio's terminology, but otherwise I'm just coopting her ideas.

more money, destroys her knowledge.

And the final group are what I'll call, a little tendentiously, the **orthodox** theories. Orthodoxy says that Anisa knew when the Battle of Agincourt was last night, since her belief satisfied every plausible criteria for testimonial knowledge. And it says she knows it today, since changing practical scenarios or interests like this doesn't affect knowledge. But it also says that the actions that Anisa and Blaise take are wrong; they are both irrational, and Blaise's is arguably immoral. And that is true because of how risky the actions are. So knowing that what you are doing is for the best is consistent with your action being faulted on epistemic grounds.

My reading of the literature is that a considerable majority of philosophers writing on these cases are orthodox. (Hence the name!) But I can't be entirely sure, because a lot of these philosophers are more vocal about opposing pragmatist views than they are about supporting any particular view. There are some views that are clearly orthodox in the sense I've described, and I really think most of the people who have opposed pragmatist treatments of these cases are orthodox, but it's possible more of them are sceptical or epistemicist than I've appreciated.

Calling this last group orthodox lets me conveniently label the other three groups as heterodox. And this lets me state what I hope to argue for in this book. I think that the interest-relative treatment of the cases is correct; and if it isn't, then at least some pragmatist treatment is correct; and if it isn't, then at least some heterodox treatment is correct.

And it's worth laying out the interest-relative case in some detail, because we can only properly assess the options holistically. Every view is going to have some very counterintuitive consequences, and we can only weigh them up when we see them all laid out.

The last claim, that every view has counterintuitive consequences, deserves some defence. I'll say much more about the challenges orthodoxy faces in section 2.3. But just to set out a simple version of the problems for each theory, observe all of the following look true.

- Sceptical theories imply that when Anisa is reading her book, she doesn't gain knowledge even though the book is reliable and she believes it because of a well-supported belief in its reliability.
- Epistemicist theories imply that Anisa and Blaise make rational choices, even though they take what look like absurd risks.
- Pragmatist theories say that offering someone a bet can cause them to lose knowledge and, presumably, that withdrawing that offer can cause them to get the knowledge back.

- Orthodox theories say that it is irrational to do something that one knows will get the best result simply because it might get a bad result.

Much of what follows in this book, like much of what's in this literature, will fall into one of two categories. Either it will be an attempt to sharpen one of these implausible consequences, so the view with that consequence looks even worse than it does now. Or it will be an attempt to dull one of them, by coming up with a version of the view that doesn't have quite as bad a consequence. Sometimes this latter task is sophistry in the bad sense; it's an attempt to make the implausible consequence of the theory harder to say, and so less of an apparent flaw on that ground alone. But sometimes it is valuable drawing of distinctions. (That is, scholasticism in the good sense.) It turns out that the alleged plausible claim is ambiguous; on one disambiguation we have really good reason to believe it is true, on another the theory in question violates it, but on no disambiguation do we get a violation of something really well-supported. I hope that they work I do here to defend the interest-relative theory is more scholastic than sophistic, but I'll leave that for others to decide.

Still, if all of the theories are implausible in one way or another, shouldn't we look for an alternative? Perhaps we should look, but we won't find any. At least if we define the theories carefully enough, the truth is guaranteed to be among them. Let's try placing theories by asking three yes/no questions.

1. Does the theory say that Anisa knew last night that the Battle of Agincourt was in 1415? If no, the theory is sceptical; if yes, go to question 2.
2. Does the theory say that Anisa is rational to play Blue-True? If yes, the theory is epistemicist; if no, go to question 3.
3. Does the theory say that Anisa still knows that the Battle of Agincourt was in 1415, at the time she chooses to play Blue-True? If no, the theory is pragmatist; if yes, the theory is orthodox.

That's it - those are your options. There are two two points of clarification that matter, but I don't think they make a huge difference.

The first point of clarification is really a reminder that these are families of views. It might be that one member of the family is considerably less implausible than other members. Indeed, I've changed my mind a fair bit about what is the best kind of pragmatist theory since I first started writing

on this. And there are a lot of possible orthodox theories. Finding out the best version of these kinds of theories, especially the last two kinds, is hard work, but it is worth doing. But I very much doubt it will lessen the implausibility of the resulting theory; some of the implausibility flows directly from how one answers the three questions.

The second point of clarification is that what I've really done here is classify what the different theories say about Anisa's case. They may say different things about other cases. A theory might take an epistemicist stand on Anisa's case, but an orthodox one on Blaise's case, for example. Or it might be orthodox about Anisa, but would be epistemicist if the blue sentence was something much more secure, such as that the Battle of Hastings was in 1066. If this taxonomy is going to be complete, it needs to say something about theories that treat different cases differently. So here is the more general taxonomy I will use.

The cases I'll quantify over have the following structure. The hero is given strong evidence for some truth p, and they believe it on the basis of that evidence. There are no defeaters, the belief is caused by the truth of the proposition in the right way, and in general all the conditions for knowledge that people worried about in the traditional (i.e., late twentieth century) epistemological literature are met. Then they are offered a choice, where one of the options will have an optimal outcome if p, but will not maximise expected value unless the probability of p is absurdly close to 1. And while hero's evidence is strong, it isn't that strong. Despite this, hero takes the risky option, using the fact that p as a key part of their reasoning. Now consider the following three questions.

1. In cases with this form, does the theory say that when the hero first forms the belief that p, they know that p? If the answer is that this is *generally* the case, then restrict attention to those cases where they do know that p, and move to question 2. Otherwise, the theory is sceptical.
2. In the cases that remain, is hero rational in taking the option that is optimal if p, but requires very high probability to maximise expected returns? If the answer is yes in *every* case, the theory is epistemicist. Otherwise, restrict attention to cases where this choice is irrational, and move to question 3.
3. In *any* of the cases that remain, does the fact that hero was offered the choice destroy their knowledge that p? If yes, the theory is pragmatic. If no, the theory is orthodox.

So I'm taking epistemicism to be a very strong theory - it says that knowledge always suffices for action that is optimal given what's known, and that offers of bets never constitute a loss of knowledge. The epistemicist can allow that the offer of a bet may cause a person to 'lose their nerve', and hence their belief that p, and hence their knowledge that p. But if they remain confident in p, they retain knowledge that p.

And I'm taking pragmatism to be a very weak theory - it says sometimes the offer of a bet can constitute a loss of knowledge. The justification for defending such a weak theory is that so many philosophers are aghast at the idea that practical considerations like this could ever be relevant to knowledge. So even showing that the existential claim is true, that sometimes practical issues matter, would be a big deal.

And orthodoxy is a weak claim on one point, and a strong claim on another. It says there are some cases where knowledge does not suffice for action - though it might take these cases to be very rare. It is common in defences of orthodoxy to say that the cases are quite rare, and use this fact to explain away intuitions that threaten orthodoxy. But it says that pragmatic factors never matter - so it can be threatened by a single case like Anisa.

2.3 Against Orthodoxy

The orthodox view of cases like Blaise's is that offering him that that does not change what he knows, but still he is irrational to take the bet. In this section, I'm going to run through a series of arguments against the orthodox view. The reason I making so many arguments is not that I think any one of them is particularly week. Rather, it is because the orthodox view is so widespread that I think we need to stress how many strange consequences it has.

2.3.1 Moore's Paradox

Start by thinking about what the Orthodox few says of rational person in Blaise's situation would do. Call this rational person Chamari. According to the orthodox view, offering someone a bit does not make them lose knowledge. So Chamari still knows when the Battle of Avignon was fought. But Chamari is rational, so Chamari will clearly decline the bet. Think about how Chamari might respond when you ask her to justify declining the bet.

> You: When was the Battle of Avignon?
> Chamari: October 25, 1415.
> You: If that's true, what will happen if you accept the bet?
> Chamari: A child will get a moment of joy.
> You: Is that a good thing?
> Chamari: Yes.
> You: So why didn't you take the bet?
> Chamari: Because it's too risky.
> You: Why is it risky?
> Chamari: Because it might lose.
> You: You mean the Battle of Avignon might not have been fought in 1415.
> Chamari: Yes.
> You: So the Battle of Avignon was fought in 1415, but it might not have been fought then?
> Chamari: Yes, the Battle of Avignon was fought in 1415, but it might not have been fought then, and that's why I'm not taking the bet.

I think Chamari has given the best possible answer at each point. But she has ended up assenting to a Moore-paradoxical sentence. In particular, she has assented to a sentence of the form *p, but it might be that not p*. And it is very widely held that sentences like this cannot be rationally assented to. Since Chamari was, by stipulation, the model for what the orthodox view thinks a rational person is, this shows that the orthodox view is false.

There are three ways out of this puzzle, and none of them seems particularly attractive.

One is to deny that there's anything wrong with where Chamari ends up. Perhaps in this case the Moore-paradoxical claim is perfectly assertable. I have some sympathy for the general idea that philosophers over-state the badness of Moore-paradoxicality (Maitra and Weatherson, 2010). But I have to say, in this instance it seems like a terrible way to end the conversation.

Another is to deny that the fact that Chamari knows something licences her in asserting it. I've assumed in the argument that if Chamari knows that *p*, she can say that *p*. But maybe that's strong an assumption. The conversation, says this reply, goes off the rails at the very first line. But on this way of thinking, it is hard to know what the point of knowledge is. If knowing something isn't sufficiently good reason to assert it, it is hard to know what would be.

The orthodox theorist has a couple of choices here, neither of them good. One is to say that although knowledge is not interest-relative, the epistemic standards for assertion are interest-relative. Basically, Chamari meets the epistemic standard for saying that *p* only if Chamari knows that *p* according to the (false!) interest-relative theory. But at this point, given how plausible it is that knowledge is closely connected with testimony, it seems we would need an excellent reason to not simply identify knowledge with this epistemic standard. The other is to say that there is some interest-invariant standard for assertion. But versions of Blaise's case show that this standard would have to be something like Cartesian certainty. So most everything we say, every single day, would be norm violating. Such a norm is not plausible.

So we get to the third way out, one that is only available to a subset of orthodox theorists. We can say that 'knows' is context-sensitive, that in Chamari's context the sentence "I know when the Battle of Agincourt was fought" is *false*, and use these facts to explain away what goes wrong in the conversation with Chamari. Armour-Garb (2011), who points out how much trouble non-contextualist orthodox theorists get into with these Moore-paradoxical claims, suggests a contextualist resolution of the puzzles. And this is probably the least bad way to handle the case, but it's worth noting just how odd it is.

It's not immediately obvious how to get from contextualism to a resolution of the puzzle. Chamari doesn't use the verb 'to know' or any of its cognates. She does use the modal 'might', and the contextualist will presumably want to say that it is context sensitive. But that doesn't look like a helpful way to solve the problem, since her assertion that the Battle might have been on a different day seems like the good part of what she says. What's problematic is the unqualified assertion about when the battle was. And we need some way of connecting contextualism about epistemic verbs to a claim about the inappropriateness of this assertion.

The standard move by contextualists here is to simply deny that there is a tight connection between knowledge and assertion (DeRose, 2002; Cohen, 2004). (So this is really a form of the response I just rejected.) What they say instead is that there is a kind of meta-linguistic standard for assertion. It is epistemically responsible to say that *p* iff it would be true to say *I know that p*. And since it would not be true for Chamari to say she knows when the Battle of Agincourt was fought, she can't responsibly say when it was fought.

The most obvious worry with this line of reasoning is with the very

idea of meta-linguistic norms like this. Imagine we were conversing with Chamari about her reasons for declining the bet in Bengali rather than English, but at every line a contribution with the same content was made? Would the reason her first answer was inappropriate be that some English sentence would be false if uttered in her context, or that some Bengali sentence would be false? If it's an English sentence, it's very weird that English would have this normative force over conversations in Bengali. If it's Bengali, then it's odd that the standard for assertion changes from language to language.

If there were a human language that didn't have a verb for knowledge, then that last point could be made with particular force. What would the contextualists say is the standard for assertion in such a language? But there is, quite surprisingly, no such language (Nagel, 2014). It's still a bit interesting to think about possible languages that do allow for assertions, but do not have a verb for knowledge. Just what the contextualists would say is the standard for assertion in such a language is a rather delicate matter.

But rather than thinking about these merely possible languages, let's return to English, and end with a variant of the conversation with Chamari. Imagine that she hasn't yet been offered the bet that Blaise is offered, and indeed that when the conversation starts, we're just spending a pleasant few minutes idly chatting about medieval history.

> You: When was the Battle of Avignon?
> Chamari: October 25, 1415.
> You: Oh that's interesting. Because you know there's this bet, and if you accept it, and the Battle of Avignon was in 1415, then a small child gets a moment of joy.
> Chamari: That's great, I should take that bet.
> You: Well, wait a second, I should tell you what happens if the Battle turns out to have been on any other date. [You explain what happens in some detail.]
> Chamari: That's awful, I shouldn't take the bet. The Battle might not have been in 1415, and it's not worth the risk.
> You: So you won't take the bet because it's too risky?
> Chamari: That's right, I won't take it because it's too risky.
> You: Why is it risky?
> Chamari: Because it might lose.
> You: You mean the Battle of Avignon might not have been fought in 1415.

> Chamari: Yes.
> You: Hang on, you just say it was fought in 1415, on October 25 to be precise.
> Chamari: That's true, I did say that.
> You: Were you wrong to have said it?
> Chamari: Probably not; it was probably right that I said it.
> You: You probably knew when the battle was, but you don't now know it?
> Chamari: No, I definitely didn't know when the battle was, but it was probably right to have said it was in 1415.

And you can probably see all sorts of ways of making Chamari's position sound terrible. The argument I'm giving here is of course just a version of the argument John ? gives arguing that contextualists have a problem with retraction. And Chamari's position does sound very bad here.

But I don't want to lean too much weight on how she sounds. Every position in this area ends up saying some strange things. The very idea that the epistemic standard for assertion could be meta-linguistic is, to my mind, even more implausible than the idea that we should end up where Chamari does.

2.3.2 Super Knowledge to the Rescue?

Let's leave Blaise and Chamari for a little and return to Anisa. The orthodox view agrees that it is irrational for Anisa to play Blue-True. So it needs to explain why this is so. I think there is a simple explanation. If she plays Red-True, she knows she will get $50; if she plays Blue-True, she does not know that - though she knows she will get at most $50. So Red-True is the weakly dominant option; she knows it won't do worse than any other option, and there is no other option that she knows won't do worse than any other option.

The orthodox theorist can't offer this explanation. They think Anisa knows that Blue-True will get $50 as well. So what can they offer instead? There are two broad kinds of explanation thet can try. First, they might offer a structurally similar explanation to the one I just gave, but with some other epistemic notion at its centre. So while Anisa knows that Blue-True will get $50, she doesn't *super-know* this, in some sense. Second, they can try to explain the asymmetry between Red-True and Blue-True in probablistic, rather than epistemic, terms. I'll discuss the first option in this subjection, and the probabilistic notion in the next subsection.

What do I mean her by *super-knows*? I mean this phrase to be a placeholder for any kind of relation stronger than knowledge that could play the right kind of role in explaining why it is irrational for Anisa to play Blue-True. So super-knowledge might be iterated knowledge. Anisa super-knows something iff she knows that she knows that … she knows it. And she super-knows that two plus two is four, but not that the Battle of Agincourt was in 1415. Or super-knowledge might be (rational) certainty. Anisa is (rationally) certain that two plus two is four, but not that the Battle of Agincourt was in 1415. Or it might be some other similar relation. My objection to this kind of response won't be sensitive to the details.

What is going to be important is that super-knowledge, whatever it is, looks like an epistemic relation. In particular, Anisa super-knows a conjunction (that she is considering) iff she super-knows each of the conjuncts. So we can't equate super-knowledge with rational credence above a threshold, because rational credence above a threshold doesn't satisfy this constraint. I'll come back to credence based explanations of Anisa's case in the next subsection.

The fact that Anisa doesn't super-know when the Battle of Agincourt was can't explain the asymmetry between Red-True and Blue-True. It can't explain why Anisa rationally must choose Red-True. Even if she super-knows that two plus two is four, she doesn't super-know the rules of the game. She has ordinary testimonial knowledge of those, just like she has ordinary testimonial knowledge about the Battle of Agincourt. In the description, I stipulated that she didn't know anything relevant about the game set up other than what I said there. So she isn't certain about the rules, and she doesn't even know that she knows them. Maybe you think that last is unrealistic, and it is important that the example is realistic. But even on a realistic treatment of the game, she won't super-know the rules. If testimony from careful historians can't generate super-knowledge, neither can testimony from game-show hosts.

In fact, her knowledge of the rules of the game, in the sense that matters, is probably weaker than her knowledge of history. It is not unknown for game shows to promise prizes, then fail to deliver them, either because of malice or incompetence. Knowledge of the game rules, in particular knowledge that she will actually get $50 if she selects a true sentence, requires some knowledge of the future. And that seems harder to obtain than knowledge of what happened in history. After all, she has to know that there won't be an alien invasion, or a giant asteroid, or an incompetent or malicious game orgAnisar.

So there is no way of understanding 'super-knows' such that 1 is true and 2 is false.

1. Anisa super-knows that if she plays Red-True, she'll win $50.
2. Anisa does not super-know that if she plays Blue-True, she'll win $50.

And that's the kind of contrast we need in order for a super-knowledge based explanation of why she should play Red-True to work.

The point I'm making here, that in thinking about these games we need to attend to the player's epistemic attitude towards the game itself, is not original Dorit Ganson (2019) uses this point for a very similar purpose, and quotes Robert Nozick (1981) making a similar point. But I've belaboured it a bit here because it is so easily overlooked. It is easy to take things that one is told about a situation, such as the rules of a game that are being played, as somehow fixed and inviolable. They aren't the kind of thing that can be questioned. But in any realistic case, that will not be how things are.

This is why I want to rest more weight on Anisa's case than on Blaise's. I can't appeal to your judgment about what a realistic version of Blaise's case would be like, because there are no realistic versions of Blaise's case. But Anisa's case is very easy to imagine and understand. And we can ask what a realistic version of it would be like. And that version would be such that the player would know what the rules of the game are, but would also know that sometimes game shows don't keep their promises, sometimes they don't describe their own games accurately, sometimes players misinterpret or misunderstand instructions, and so on. This shouldn't lead us to scepticism: Anisa knows what game she's playing. But she doesn't super-know what game she's playing. And that means she doesn't super-know that she'll win if she plays Red-True.

2.3.3 Rational Credences to the Rescue?

So imagine the orthodox theorist drops super-knowledge, and looks somewhere else. A natural alternative is to use credences. Assume that the probability that the rules of the game are as described is independent of the probabilities of the red and blue sentence. And assume that Anisa must, if she is to be rational, maximise expected utility. Then we get the natural result that Anisa should pick the sentence that is more probably

true.[2] And that can explain why she must choose Red-True, which is what the orthodox theorist needed to explain.

This kind of approach doesn't really have any place for knowledge in its theory of action. One should simply maximise expected utility; since doing what one knows to be best might not maximise expected utility, we shouldn't think knowledge has any particularly special role.

But there are many problems with this kind of approach. Several of these will be discussed elsewhere in this book at more length, so I'll here just point to those other discussions. But some others I'll address directly.

Like the view discussed in subsection 2.3.1 that separates knowledge from assertion, separating knowledge from action leads to strange consequences. As ? points out, once we break apart knowledge from action in this way, it becomes hard to see the point of action. And it's worth pausing a bit more over the bizarreness of the claim that Blaise knows that taking the bet will work out for the best, but he shouldn't take it - because of it's possible consequences!

If one excludes knowledge from having an important role in one's theory of decision, one ends up having a hard time explaining how dominance reasoning works. But it is a compulsory question for a theory of decision to explain how dominance reasoning works. Among other things, we need a good account of how dominance reasoning works in order to handle Newcomb problems, and we need to handle Newcomb problems in order to even clearly state a theory of expected utility maximisation. That little argument was very compressed, but I'll return frequently through the book to issues about dominance reasoning, and for now I think it's enough to leave it with this sketch.

Probabilistic models of reasoning and decision have their limits. But what we need to explain about the Red-Blue game goes beyond those limits. So probabilistic models can't be the full story about the Red-Blue game.

To see this, imagine for a second that the Blue sentence is not about the Battle of Agincourt, but is instead a slightly more complicated arithmetic truth, like *Thirteen times seventeen equals two hundred and twenty one*, or a slightly complicated logical truth, like $\neg q \rightarrow ((p \rightarrow q) \rightarrow \neg p)$. If either of those are the blue sentence, then it is still uniquely rational to play Red-True. But the probability of each of those sentences is one. So rational

[2] Strictly speaking, we need one more assumption - namely that for any unexpected way for the game to be, the probability of it being that way is independent of the truth of both the red and blue sentences. But it's natural to assume that this is true.

choice is more demanding than expected utility maximisation. In sections 6.2 and 6.3 I'll go over more cases of propositions whose probability is 1, but which should be treated as uncertain even it is certain that two plus two is four. The lesson is that we can't just use expected utility maximisation to explain the Red-Blue game.

Finally, we need to understand the notion of probability that's being appealed to in this explanation. It can't be some purely subjective notion, like credence, because that couldn't explain why some decisions are rational and others aren't. And it can't be some purely physical notion, like chance or frequency, because that won't even get the cases right. (What is the chance, or frequency, of the Battle of Agincourt being in 1415?) It needs to be something like evidential probability. And that will run into problems in versions of the Red-Blue game where the Blue sentence is arguably (but not certainly) part of the player's evidence. I'll end my discussion of orthodoxy with a discussion of cases like these.

2.3.4 Evidential Probability

No matter which of these explanations the orthodox theorist goes for, they need a notion of evidence to support them. Let's assume that we can find some doxastic attitude *D* such that Anisa can't rationally stand in *D* to *Play Blue-True*, and that this is why she can't rationally play Blue-True. Then we need to ask the further question, why doesn't she stand in relation *D* to *Play Blue-True*? And presumably the answer will be that she lacks sufficient evidence. If she had optimal evidence about when the Battle of Agincourt was, she could play Blue-True, after all.

The orthodox theorist also has to have an interest-invariant account of evidence. It's logically possible to have evidence be interest-relative, but knowledge be interest-neutral, but it is very hard to see how one would motivate such a position.

And now we run into a problem. Imagine a version of the Red-Blue game where the blue sentence is something that, if known, is part of the player's evidence. If it is still irrational to play Blue-True, then any orthodox explanation that relies on evidence sensitive notions (like superknowledge or evidential probability) will be in trouble. The aim of this subsection is to spell out why this is.

So let's imagine a new player for the red-blue game. Call her Parveen. She is playing the game in a restaurant. It is near her apartment in Ann Arbor, Michigan. Just before the game starts, she notices an old friend, Rahul, across the room. Rahul is someone she knows well, and can ordi-

narily recognise, but she had no idea he was in town. She thought Rahul was living in Italy. Still, we would ordinarily say that she now knows Rahul is in town; indeed that he is in the restaurant. As evidence for this, note that it would be perfectly acceptable for her to say to someone else, "I saw Rahul here". Now the game starts.

- The red sentence is *Two plus two equals four*.
- The blue sentence is *Rahul is in this restaurant*.

And here is the problem. On the one hand, there is only one rational play for Parveen: Red-True. She hasn't seen Rahul in ages, and she thought he was in Italy. A glimpse of him across a crowded restaurant isn't enough for her to think that *Rahul is in this restaurant* is as likely as *Two plus two equals four*. She might be wrong about Rahul, so she should take the sure money and play Red-True. So playing the red-blue game with these sentences makes it the case that Parveen doesn't know where Rahul is. This is another case where knowledge is interest-relative, and at first glance it doesn't look very different to the other cases we've seen.

But take a second look at the story for why Parveen doesn't know where Rahul is. It can't be just that her evidence makes it certain that two plus two equals four, but not certain that Rahul is in the restaurant. At least, it can't be that unless it is not part of her evidence that Rahul is in the restaurant. And if evidence is not interest-relative, then I think we should say that it is part of Parveen's evidence that Rahul is in the restaurant. This isn't something she infers; it is a fact about the world she simply appreciated. Ordinarily, it is a starting point for her later deliberations, such as deliberations about whether to walk over to another part of the restaurant to say hi to Rahul. That is, ordinarily it is part of her evidence.

So the orthodox theorist has a challenge. If they say that it is part of Parveen's evidence that Rahul is in the restaurant, then they can't turn around and say that the evidential probability that he is in the restaurant is insufficiently high for her to play Blue-True. After all, it's evidential probability is one. If they say that it is no part of Parveen's evidence that Rahul is in the restaurant because she is playing this version of the Red-Blue game, they give up orthodoxy. So they have to say that our evidence never includes things like Rahul is in the restaurant.

This can be generalised. Take any proposition such that if the red sentence was that two plus two is four and that proposition was the content of the blue sentence, then it would be irrational to play Blue-True. Any orthodox explanation of the Red-Blue game entails that this proposition

is no part of your evidence - whether you are playing the game or not. But once we strip all these propositions out of your evidence, you don't have enough evidence to rationally believe, or even rationally make probable, very much at all.

Descartes, via very different means, walked into a version of this problem. And his answer was to (implicitly) take us to be infallible observers of our own minds, and (explicitly) offer a theistic explanation for how we can know about the external world given just this psychologistic evidence. Nowadays, most people think that's wrong on both counts: we can be rationally uncertain about even our own minds, and there is no good path from purely psychological evidence to knowledge of the external world. The orthodox theorist ends up in a state worse than Cartesian scepticism.

2.4 Odds and Stakes

If orthodox views are wrong, then it is important to get clear on which heterodox view is most plausible.[3] I'm defending a version of the pragmatic view. But it's a different version to the most prominent versions defended in the literature. The difference can be most readily seen by looking at the class of cases that have motivated pragmatic views.

The cases involve a subject making a practical decision. The subject has a safe choice, which has a guaranteed return of S. And they have a risky choice. If things go well, the return of the risky choice is $S + G$, so they will gain G from taking the risk. If things go badly, the return of the risky choice is $S - L$, so they will lose L from taking the risk. What it takes for things to go well is that a particular proposition p is true. All of this is known by the subject facing the choice. What the subject doesn't (uncontroversially) know is that they satisfy all the conditions for knowing p that would have been endorsed by a well-informed epistemologist circa 1997. (That is, by a proponent of the traditional view.) So p is true, and things won't go badly for them if they take the risk. But still, in a lot of these cases, there is a strong intuition that they do not know that p, and as I've just been arguing, that is hard to square with the idea that they know that p. So assuming the traditional view is right about the subject as they were before facing the practical choice, having this choice in front of them causes them to lose knowledge that p.

But what is it about these choices that triggers a loss of knowledge? There is a familiar answer to this, one explicitly endorsed by Hawthorne

[3] This section is based on §3 of my (?).

(2004) and ʔ. It is that they are facing a 'high stakes' choice. Now what it is for a choice to be high stakes is never made entirely clear, and Anderson and Hawthorne (2019a) show that it is hard to provide an adequate definition in full generality. But in the simple cases described in the previous paragraph, it is easy enough to say what a high stakes case is. It just means that L is large. So one gets the suggestion that practical factors kick in when faced with a case where there is a chance of a large loss.

This is not the view I defend. I think L matters, but only indirectly. What is (typically) true in these cases is that the subject should maximise expected utility relative to what they know.[4] And taking the risky choice maximises expected utility only if this equation is true.

$$\frac{\Pr(p)}{1-\Pr(p)} > \frac{L}{G}$$

The left hand side expresses the odds that p is true. The right hand side expresses how high those odds have to be before the risk is worth taking. If the equation fails to hold, then the risk is not worth taking. And if risk is not worth taking, then the subject doesn't know that p

Since the numerator of the right hand side is L, then one way to destroy knowledge that p is to present the subject with a situation where L is very high. But it isn't the only way. Since the denominator of the right hand side is G, another way to destroy knowledge that p is to present the subject with a situation where G is very low.

In effect, we've seen such a situation with Anisa. But to make the parallel to Blaise's case even clearer, consider Darja's case. She has been reading books about World War One, and yesterday read that Franz Ferdinand was assassinated on St Vitus's Day, June 28, 1914. She is now offered a chance to play a slightly unusual quiz game. She has to answer the question *What was the date of Franz Ferdinand's assassination?* If she gets it right, she wins $50. If she gets it wrong, she wins nothing. Here's what is strange about the game. She is allowed to Google the answer before answering. So here are the two live options for Darja. In the table, and in what follows, p is the proposition that Franz Ferdinand was indeed assassinated on June 28, 1914.

	p	$\neg p$

[4] This simplifies a little the relationship between rational choice and expected utility maximisation. Later in the book I'll have to be much more careful about this relationship. See chapter 7 for many more details.

Say "June 28, 1914"	50	0
Google the answer	$50 - \varepsilon$	$50 - \varepsilon$

If Darja has her phone near her, and has cheap easy access to Google, then ε might be really low. And then she should take the safe option, unless she is incredibly sure that the book she read is reliable, and that she has precisely remembered it. In a lot of realistic cases, that won't happen, and $\Pr(p)$ will be too low for her to take the risky option of saying "June 28, 1914". She should take the safe option of Googling the answer. And that means she doesn't know that p, even if she remembers reading it in a book that is actually reliable. Facing a long odds bet can cause knowledge loss, even in low stakes situations.

I'm not the first to focus on these long odds/low stakes cases. Jessica Brown (2008, 176) notes that these cases raise problems for the stakes-centric version of IRT. And Anderson and Hawthorne (2019a) argue that once we get beyond the simple two-state/two-option choices, it isn't at all easy to say what situations are and are not high-stakes choices. I'll return to their objection in chapter 8, but for now I just want to note that the version of IRT I'm defending doesn't give any special significance to high stakes choices. What makes knowledge hard is, to a first approximation, facing a long odds bet, not facing a high stakes bet.

2.5 Theoretical Interests Matter

When saying why I called my theory IRT, one of the reasons I gave was that I wanted theoretical, and not just practical, interests to matter to knowledge.[5] This is also something of a break with the existing literature. After all, Jason Stanley's book on interest-relative epistemology is called *Knowledge and Practical Interests*. And he defends a theory on which what an agent knows depends on the practical questions they face. But there are strong reasons to think that theoretical reasons matter as well.

In section 2.4, I suggested that someone knows that p only if the rational choice to make would also be rational given p. That is, someone knows that p only if the answer to the question *What should I do?* is the same unconditionally as it is conditional on p. My preferred version of IRT generalises this approach. Someone knows that p only if the rational answer to a question she is interested in is the same unconditionally as it

[5] This section is based on §4 of my (?).

is conditional on *p*. Interests matter because they determine just what it is for the person to be interested in a question. If that's how one thinks of IRT, the question of this section becomes, should we restrict questions the agent is interested in to just being questions about what choice to make? Or should they include questions that turn on her thoeretical interests, but which are irrelevant to choices before her? There are two primary motivations for allowing theoretical interests as well as practical interests to matter.

The first comes from the arguments for what Jeremy Fantl and Matthew McGrath call the Unity Thesis (Fantl and McGrath, 2009, 73–6). They are interested in the thesis that whether or not *p* is a reason for someone is independent of whether they are engaged in practical or theoretical deliberation. But one doesn't have to be so invested in the ideology of reasons to appreciate their argument. Note that if only practical interests matter, then they know different things when considering the question *What to do in situation* S in situation *S* and other situations. And if they know different things, those differential pieces of knowledge could lead to different answers. And that's very unintuitive. After all, they might be deliberating about this question because situation *S* might arise, and they want to be practically ready for it.

Let's make that a little less abstract. Imagine Anisa is not actually faced with the choice between Red-True, Blue-True, Red False and Blue-False with these particular red and blue sentences. In fact, she has no practical decision to make that turns on the date of the Battle of Agincourt. But she is idly musing over what she would do if she were playing that game. If she knows when the battle was, then she should be indifferent between Red-True and Blue-True. After all, she knows they will both win $50. But intuitively she should think Red-True is preferable, even in the abstract setting. And this seems to be the totally general case.

The general lesson is that if whether one can take *p* for granted is relevant to the choice between A and B, it is similarly relevant to the theoretical question of whether one would choose A or B, given a choice. And since those questions should receive the same answer, if *p* can't be known while making the practical deliberation between A and B, it can't be known while musing on whether A or B is more choiceworthy.

There is a second reason for including theoretical interests in what's relevant to knowledge. There is something odd about the following reasoning: The probability of *p is precisely x*, therefore *p*, in any case where $x < 1$. It is a little hard to say, though, why this is problematic, since we

often take ourselves to know things on what we would admit, if pushed, are purely probabilistic grounds. The version of IRT that includes theoretical interests allows for this. If we are consciously thinking about whether the probability of p is x, then that's a relevant question to us. Conditional on p, the answer to that question is clearly no, since conditional on p, the probability of p is 1. So anyone who is thinking about the precise probability of p, and not thinking it is 1, is not in a position to know p. And that's why it is wrong, when thinking about p's probability, to infer p from its high probability.

Putting the ideas so far together, we get the following picture of how interests matter. Someone knows that p only if the evidential probability of p is close enough to certainty for all the purposes that are relevant, given their theoretical and practical interests. Assuming the background theory of knowledge is non-sceptical, this will entail that interests matter.

2.6 Global Interest Relativity

IRT was introduced as a thesis about knowledge. I'm going to argue in chapter 6 that it also extends to rational belief. Not every case where interests matter to knowledge generates a Dharmottara case. But we need not stop there. At the extreme, we could argue that every epistemologially interesting notion is interest-relative. Doing so gives us a global version of IRT. And that is what I'm going to defend here.

Jason Stanley (2005) comes close to defending a global version. He notes that if one has both IRT, and a 'knowledge first' epistemology (Williamson, 2000), then one is a long way to towards global IRT. Even if one doesn't accept the whole knowledge first package, but just accepts the thesis that evidence is all and only what one knows, then one is a long way towards globalism. After all, if evidence is interest-relative, then probability, justification, rationality, and evidential support are interest-relative too.

That's close to the path I'll take to global IRT, but not exactly it. In chapter 5 I'm going to argue that evidence is indeed interest-relative, and so all those other notions are interest-relative too. But the version of IRT I'll put forward implies that evidence is a subset of knowledge.

There is a deep puzzle here for IRT that for a long time I couldn't see a way out of. On the one hand, the arguments for IRT look like they will generalise to arguments for the interest-relativity of evidence. (This is something Tom Donaldson convinced me of in conversations some years

back.) On the other hand, the explanation I want to offer of cases like Anisa's presupposes that we can identify Anisa's evidence independent of her interests. I want to say that Anisa shouldn't play Blue-True because the evidential probability of the blue sentence being true is lower than the evidential probability of the red sentence being true. And she can't know the blue sentence is true because she can't play Blue-True. This turns into a nice story about when changes of interests lead to changes in belief if we can independently identify Anisa's evidence, and hence the evidential probability of different propositions. But the story looks much less nice if interests also affect evidence.

The aim of chapter 5 is to tell a story that avoids the worst of these problems. On the story I'll tell, evidence is indeed interest-relative. And that means we can't tell a simple story about precisely when changes in interests will lead to changes in knowledge. But it will still be true that people lose knowledge when the evidential probability of a proposition is no longer high enough for them to take it for granted with respect to every question they are interested in. And I will be able to say how interests impact evidence in a way that doesn't require antecedently identifying how interests impact knowledge, so the story will still be somewhat reductive. But it won't be as simple a story as I once believed in.

2.7 Neutrality

This book defends, at some length, the idea that knowledge is interest-relative. But I'm staying neutral on a number of other topics in the vicinity.

Most notably, I'm not taking any stand on whether *contextualist* theories of knowledge are true or false. If you think that contextualism is true, then what I'm defending is that the view that 'knowledge' picks out in this context, and in most other contexts, is interest-relative.

Contextualist theories of knowledge have a lot in common with interest-relative theories. The kind of cases that motivate the interest-relative theories, cases like Anisa's and Blaise's, also motivate contextualism. They might even be seen as competitors, since they are offering rival explanations of similar phenomena. But they are not strictly inconsistent. Consider principles A and B below.

A. A's utterance that *B knows that p* is true only if for any question *Q?* in which A is interested, the rational answer for B to give is the same unconditionally as it is conditional on *p*.

B. A's utterance that *B knows that p* is true only if for any question *Q?* in which B is interested, the rational answer for B to give is the same unconditionally as it is conditional on *p*.

I endorse principle B, and that's why I endorse an interest-relative theory of knowledge. If I endorsed principle A, then I would be (more or less) committed to a contextualist theory of knowledge. And principle A is not inconsistent with principle B.[6]

It isn't hard to see why cases like Anisa and Blaise can move one to endorse principle A, and hence contextualism. It would be very odd for Anisa to say "This morning, I knew the Battle of Agincourt was in 1415." That's odd because she can't now take it as given that the Battle of Agincourt was in 1415, and in some sense she wasn't in any better or worse evidential position this morning with respect to the date of the battle. Perhaps, and this is the key point, it would even be false for Anisa to say this now. The contextualist, especially the contextualist who endorses principle A, has a good explanation for why that's false. The interest-relative theorist doesn't have anything to say about that. Personally I think it's not obvious whether this would be false for Anisa to say, or merely inappropriate, and even if it is false, there may be decent explanations of this that are not contextualist. But there is clearly an argument for contextualism here. And it isn't one that I'm going to endorse or reject.

As I've already noted, I'm making heavy use of the principle that Jessica Brown calls K-Suff. I'm going to defend that at much greater length in chapter ??. What I'm not defending is the converse of that principle, what she calls K-Nec.

K-Nec An agent can properly use *p* as a reason for action only if she knows that *p*.

The existing arguments for and against K-Nec are intricate and interesting, and I don't have anything useful to add to them. All I will note is that the argument of this book doesn't rely on K-Nec, and I'm just going to set it aside.

And I'm obviously not going to offer anything like a full theory of knowledge. I am defending a particular necessary condition on knowledge. That condition, plus some commonsensical claims about what we know in

[6]There is a technical difficulty in how to understand one person answering an infinitival question that another person is asking themselves. But the points I'm making in this section aren't sensitive to this level of technical detail.

ordinary situations, entails that knowledge is interest-relative. And that's just about as far as I'll go.

I will be making one claim about how interests typically enter into the theory of knowledge. I'll argue that there is a certain kind of defeater. A person only knows that *p* if the belief that *p* coheres in the right way with the rest of their attitudes. What's 'the right way'? That, I argue, is interest-relative. In particular, some kinds of incoherence are compatible with knowledge if the incoherence concerns questions that are not interesting.

So the impact of interests is (typically) very indirect. Even if the other conditions for knowledge are satisifed, someone might fail to know something because it doesn't cohere well with the rest of their beliefs. But there is an exception to this exception clause. Incoherence with respect to uninteresting questions is compatible with knowledge.

This is going to matter because it affects how we think about what happen when interests change. It is odd to think that a change in interests could make one know something. But it isn't as odd to think that a change in interests could block or defeat something that was potentially going to block or defeat an otherwise well supported belief from being knowledge. This is something I will return to repeatedly in chapter 9.

Chapter 3

Belief

3.1 Beliefs and Interests

In my earliest work on interest-relativity, the 2005 paper "Can We Do Without Pragmatic Encroachment", I argued for a belief-first approach to interest-relativity. In particular, I conceived of that paper as the start of a project where I would argue (a) that belief itself is interest-relative, and (b) that all the interest-relativity in epistemology could trace back to the interest-relativity of belief. The core idea was that the metaphysics of mind should be interest-relative, but the normative theory of mind need not be. Of course, if a state is by its nature interest-relative, then whether one ought be in that state could quite easily turn out to be interest-relative as well. But the thought was that interest-invariant norms, combined with interest-relative metaphysics, could explain all the phenomena.

This approach did not work out. The model of that paper assumed a very high degree of rationality on the part of the agents being modelled, and it turns out to be very hard to extend the model to less rational agents. (Kieran Setiya was the first to point out this problem to me.) Jacob Ross and Mark Schroeder (2014) pointed out that the way the model handled propositions that are irrelevant to one's practical interests wasn't working. In "Games, Beliefs and Credences" I suggested a 'fix', but I didn't really appreciate how much the fix undermined the motivations for the original view.

In this book I'm taking a different approach. For one thing, I am defending a much more expansive picture of how interests affect epistemology. In chapter 5 I'm going to argue that evidence itself is interest-relative, and hence so is everything that depends on evidence. And in

chapter 4, I'm going to argue that a version of what Jason Stanley (2005) calls 'ignorant high stakes' cases shows that some of the interest-relativity of knowledge is not tracable to the interest-relativity of either belief or of rational belief.

And I'm going to offer a much more complicated story about the constitutive connection between interests and beliefs than I offered earlier. While I still think there is scuh a connection, I'm much more sympathetic to the suggestion by Jennifer Nagel (2008; 2010) that often there is a merely causal connection. Some people's beliefs change when the practical situation changes because the situation causes them to see that it would be a mistake to hold on to their prior beliefs. This isn't interest-relativity in the sense we're most interested in. Where I disagree with Nagel is that I think that for at least some people, the change in belief upon change in interests is more direct, and plausibly the change in interests is partially constitutive of the change in belief. But the situation is much messier than I had realised in earlier work, and Nagel's work revealed how I'd been oversimplifying things.

The positive theory I'm going to develop owes a lot to proposals by Dorit Ganson (2008; 2019). Like her, I'm going to develop a theory where we first say what it is to have a belief in normal cases, then include an exception clause for what happens in special cases, such as high-stakes or long-odds. The details will differ in some respects, but the underlying architecture will be the same.

But perhaps the biggest difference is one motivated by work by Jonathan Weisberg (2013; 2020). In the 2005 paper, I argued that what we should say someone believed is something we reconstruct from their patterns of preferences. At a high level of generality, my theory was that you could look at the outputs of someone's deliberation, and construct from that what they believed. This was a quite radical version of radical interpretation. I now think I was looking in the wrong spot. What someone believes is a matter of what inputs to deliberation they are willing to accept, not the results of those deliberations. If we assume perfect rationality, this distinction may not matter. The outputs of deliberations (taken collectively) might well imply what inputs had been accepted. But that's too strong an assumption; and for anything other than perfectly rational thinkers, we can't infer what starting points they are willing to accept from what points they end up at.

What's essential to belief, I now think, is that to believe something is to be willing to use it as a starting point in deliberation. That slogan needs

a lot of qualification to be a theory, but as a slogan it isn't a bad starting point.

3.2 Maps and Legends

Beliefs, Frank Ramsey famously said, are maps by which we steer (Ramsey, 1990, 146). I agree, and I think you can turn this into an argument that belief should be interest-relative as well. The end of the argument is a little tricky, and there is a rival position that I don't have a particularly strong objection to. But let's first see why the map based picture of belief naturally leads to interest-relativity.

When I was growing up in car-dependent, suburban Melbourne, the main street directory that was used was the Melways. This was a several hundred page thick book that most people kept a copy of in their car. It largely consisted of page after page of 1:20,000 scale maps of the Melbourne suburbs, plus more detailed maps of the inner city, and then progressively less detailed maps of the rural areas around Melbourne, the rest of Victoria, and finally of the rest of the country. And it was everywhere. It was common for store advertisements, party invitations and event announcements to include the Melways page and grid coordinates of the location. In fact I was a little shocked when I moved to America and I found it was socially expected (in those pre-Google Maps days) that you would give people something like turn by turn directions to a location. I was used to just telling people where something was, i.e., giving them the Melways grid coordinates, and letting them use the map to get themsleves there. The Melways really was, collectively, the map by which we steered.

But you wouldn't want to use it for everything. You wouldn't want to use it as a hiking map, for example. For one thing, it was much too heavy. For another, it was patchy on which walking trails it even included, and had almost no usable topographical information. You steer yourself by one map when you drive, and another map (or set of maps) when you hike. What one steers by is a function of one's interests. And the same, I think, is true of belief. Beliefs are interest relative because to believe something is to steer yourself by a map that represents the world as being that way, and which map one will steer by is sensitive to one's interests.

Maybe you think this argument leans too heavily on Ramsey's analogy of beliefs and maps. But once you see the structure of the case, you can get more purely cognitive examples. (And this in turns helps us see the brilliance of Ramsey's metaphor.) Think of the role that simple economic

models play in our thought. For a lot of purposes, I'll use a simple supply-demand model to work out the effects of a change in market conditions. But if a lot rides on the matter, or if I'm trying to work out something that changes a lot of markets at once (such as a rise in the minimum wage), I won't take these models as anything more than vaguely indicative of the truth. When trying to steer through tricky economic waters, which models I steer by will be a function of what I'm interested in at the time. Still, let's not overstate this. If you ask me what effect a new tax on widgets will have, I'll conclude it will raise the price of widgets, but by slightly less than the amount of the tax, and decrease the quantity of widgets bought, and that the quantity of those changes will be a function of the slopes of the supply curve and demand curve for widgets. That conclusion is a belief, notwithstanding any hesitation I have about using that very model in more complicated cases. I believe what simple models say, but only when there are no practical or theoretical reasons to override the model.

So it looks like belief is interest-relative, and that's for deep reasons about the role that belief plays in our cognitive economy. But note that things get complicated when we stop focussing on what normal (or normal-ish) people do, and think about less common reactions. After all, a metaphysics of belief should have something to say about them too.

Consider a person, call them Stubbie, who uses the same maps and models for every task. They use the Melways for hiking, they make macroeconomic forecasts using simple supply-demand models, and so on. And they do this even though they know full well that there are excellent reasons to be more flexible. What should we say about Stubbie?

I think we should say that Stubbie is irrationally stubborn, and part of his irrationality consists in steering by the same map, in holding onto the same beliefs, in situations where this is uncalled for. Stubbie really believes that a simple supply-demand model is predictive even in complicated cases. He's wrong, and his evidence shows that he is wrong, but our theory of belief had better allow for some false, irrational beliefs.

Stubbie's example shows that while one's beliefs should be interest-relative, they need not be. One should steer by a map suitable to the circumstances. But if one stubbornly steers by the same map come what may, the fact that it would be advisable to steer by different maps at different times does not affect what one believes. Stubbie really is steering himself by the Melways when hiking, and he really believes the simple economic model he uses.

We have already seen another pair of cases that are like Stubbie in this

respect - Anisa and Blaise. Both of them are irrational. And their irrationality consists in holding onto beliefs that they should abandon. But it is a mistake to say that because their interests have changed, that makes their beliefs change. What's true is that because their interests changed, their beliefs should have changed, but weirdly they did not. Interests change normative requirements; they don't on their own make one satisfy those changed requirements.

This refutes a position that I never explicitly adopted, but which I was I think implicitly committed to: that it is part of having beliefs that those beliefs are interest-relative. That isn't right. Some people are like Stubbie, and have interest-invariant beliefs. You shouldn't be like that; it's really irrational. But we can still take the intensional stance towards Stubbie.

There is a variant of Stubbie's case that raises particularly hard questions for my view. Imagine that Stubbie is disposed to keep taking the simple model for granted. But now he is faced with a decision where the simple model has to be exactly right for its predictions to be accurate, and the costs of being wrong are enormous. And further imagine that Stubbie sees all this, and the shock of having that much at stake causes him to reconsider. So he drops the simple supply-demand model.

This is not, I think, a case of interest-relativity of belief. Rather, it is like the kind of case Jennifer Nagel (2010) discusses, when she talks about beliefs being causally sensitive to interests. And this shows we have to be careful, more careful than I was in earlier work, to be sure that a case of interest-sensitivity is really a case where belief is constitutively, and not merely causally, sensitive to interests.

In cases of belief formation, it might be very hard to tease these two apart. Imagine that Blaise has a sister, Deja, who has read the same book about the Battle of Agincourt. And she is asked to bet on the proposition *p*, which is *That the Battle of Agincourt took place 374 years (to the nearest year) before the storming of the Bastille*. The odds are the same as in Blaise's case: a moment's joy for an infant if she's right; The Bad Place for all of us if she's wrong. She declines the bet, wisely thinking it isn't worth the risk. She is more certain that the Bastille was stormed in 1789 than she is that these are the terms of the bet, and she can subtract 1415 from 1789, so her declining is solely down to her worries about when the Battle of Agincourt was.

But note that while her beliefs before being offered the bet entailed that *p* was true, she didn't actually believe it. She hadn't given any thought to this kind of comparative claim. She was disposed to believe it upon

considering the matter, but that's not sufficient for actual belief. And, when faced with the proposition in precisely this context, she doesn't form the belief.

Here we get to a philosophical question that I have little idea how to answer. Deja doesn't believe *p*. And the explanation for this non-belief in part goes via her interests in keeping humanity out of The Bad Place. But is this a case where Deja's beliefs are causally, or constitutively, sensitive to her interests? I don't really know, and I don't really know how to motivate one view or another. If this was the only kind of case where interests mattered to beliefs, arguably the most plausible, and most conservative, way to treat the case would be to say that Deja's interests have a merely causal impact on her beliefs.

But this isn't the only kind of case to consider, and it isn't the one that I think is strongest for the view that interests can be constitutively relevant to beliefs. To see this, we need to add one last sibling to Blaise and Deja's family, their sister Eulalie. She has read the same book, and oddly last night she was also asked whether she wanted to bet on *p*. But in her case the odds were more favorable. Should she take the bet, then she would win €10 if *p* is true, and lose €10 if it is false. She took the bet and (as best she can remember), won the bet.

Now Eulalie is offered the same bet that Deja is offered. She declines it, and she was always disposed to decline any such bet. She has no inclination whatsoever to risk the future of humanity on something she read in a history book. She doesn't believe *p*, and this non-belief is explained by her interests. Is this a case where interests are causally or constitutively relevant to belief? I think you can make a case either way, but the case for it being a constiutive connection is a little stronger.

Because not a lot turns on this, and because I'm not that sure myself what the right answer is, I'll just note a handful of facts about Eulalie's case. As I've described the case, the high stakes bet didn't have to trigger anything like anxiety in her. All that was required to have her lose the belief was an appreciation of the situation she faced, and the automatic activity of some cognitive processes. That sounds more like a case where interests are constitutively relevant than one where they cause something that blocks belief. Nagel's positive argument that interests normally matter because they cause something like anxiety that blocks belief largely focusses on cases like Deja, who is forming a belief, not on cases like Eulalie, who merely loses one. And, perhaps most importantly, it is easier to formulate a positive theory of belief, one that covers both tacit and explicit beliefs

about relevant and irrelevant matters, if one takes Eulalie's example to be one where interests play a constitutuve role. I'll now turn to building just that theory.

3.3 Taking As Given

To start towards a positive theory of belief, it helps to think about the following example, featuring a guy I'll call Sully. (This example is going to resemble the examples involving Renzo in Ross and Schroeder (2014), and at least for a while, my conclusions are going to resemble theirs as well.) Sully is a fan of the Boston Red Sox, and one of the happiest days of his life was when the Red Sox broke their 86 year long curse, and won baseball's World Series in 2004. He knows, and hence believes, that the Red Sox won the World Series in 2004. He likes their chances to win again this year, because in Sully's heart, hope always springs eternal.

It's now the start of a new baseball season, and Sully is offered, for free, a choice between the following two bets.

- Bet A wins $50 if the Red Sox win the World Series this year, and nothing otherwise.
- Bet B wins $60 if the same team wins the World Series this year as won in 2004, and nothing otherwise.

For Sully, this choice is a no-brainer. If the Red Sox win this year, he wins more money taking B than A. If the Red Sox don't win this year, he gets nothing either way. So it's better to take B than A, and that's what he does.

What Sully has done here is use dominance reasoning, in particular weak dominance reasoning. One option weakly dominates another if it might have a better return, and can't have a worse return. Weak dominance is used as an analytical tool in game theory. But it is also a form of inference that non-theorists, like Sully can use. (Though unless they've taken a game theory course they might not use this phrase to describe it.)

Sully's case can be distinguished from that of his more anxious friend Mack. Mack is also a big Red Sox fan, and also looks back on that curse-busting World Series win with fondness. But if you offer Mack the choice between these two bets, he'll hesitate a bit. He'll wonder if he's really sure it was 2004 that the Red Sox won. Maybe it was 2005 he thinks. He'll eventually think that even if he's not completely sure that it was 2004, it

was very likely 2004, and so it is very likely that bet B will do better, and that's what he will take.

Even if Sully and Mack end up at the same point, they have used very different forms of reasoning. Sully uses weak dominance reasoning, while Mack uses probabilistic reasoning. Sully takes the fact that the Red Sox won in 2004 as given, while Mack just takes it to be very likely. The big thing I want to rely on here is that these are very different psychological processes. Neither of these guys is doing something that approximates, or simplifies, the other; they both take bet B, but they get to that conclusion via very different routes.

There is a theoretical analog to this psychological point. Most game theorists think that weak dominance reasoning can be iterated more or less indefinitely. (Though there are some notable exceptions.) But few think that likelihood reasoning can be iterated indefinitely. This reflects the fact that they are very different kinds of reasoning. Dominance reasoning is pre-probabilistic.

Sully's reasoning isn't just dominance reasoning. It's dominance reasoning that relies on some very specific assumptions. When Sully reasons that A can't do better than B, he's not drawing any kind of logical or metaphysical point. It's logically and metaphysically possible that the Red Sox lost in 2004. For that matter, and this is a point Ganson (2019) stresses, it's logically and metaphysically possible that the payouts for A and B are other than what Sully thinks they are.

And thought he might not make it explicit, at some level Sully surely knows this. If pushed, he'd endorse the conditional "If I've misremembered when the curse-busting World Series win was, and the Red Sox didn't win in 2004, then bet A might do better than bet B". So while he is disposed to use dominance reasoning in deciding whether to take A or B, this disposition rests on taking some facts about the world for granted.

Recall the disjunctive way that Sully reasoned. Either the Red Sox will win this year or they won't. Either way, I won't do better taking bet A, but I might do better taking bet B. So I'll take bet B. This reasoning - not just the reasons Sully has but his reasoning - can be appropriately represented by the kind of decision table that is familiar from decision theory or game theory.

	Red Sox Win	Red Sox Don't Win
Take Bet A	$50	$0
Take Bet B	$60	$0

Focus for now on the columns in this table. Sully takes two possibilities seriously: that the Red Sox win this year, and that they don't. The 'possibilities' here are possibilities in the sense of Humberstone (1981). They have content - in one of them the Red Sox win, in the other they don't, but they don't settle all facts. In the right-hand column, there is no fact of the matter about which team wins the World Series. In neither column is there a fact of the matter about what Sully will have for lunch tomorrow. If you want to think of these in terms of worlds, they are both very large sets of worlds, and within those sets there is a lot of variability.[1]

But there is more to the content of each column than what is explicity represented in the header row. In each column, for example, the Red Sox won in 2004. That's why Sully can put those monetary payoffs into the cells. And in each column, the terms of the bet are as Sully knows that they are. In sets of worlds terms, the sets that are represented by the columns are exclusive, but far from exhaustive.

Consider those propositions which are true according to all of the columns in this table. Say a proposition is *taken as given* in a decision problem when it the decider treats one option as dominating another, and does so in virtue of a table in which that proposition is true in every column. Then here is one principle about belief that seems to be very plausible.

Given S believes that p only if there is some possible decision problem such that S is disposed to take p as given when faced with that problen.

Given is logically weak in one respect, and strong in another. It only requires that S be willing to take p for granted in one possible choice. It doesn't have to be a likely, or even particularly realistic choice. Sully is unlikely to have strangers offer him these free money bets. But given how representationally sparse decision tables are, for something to be true in all columns of a decision table is a very strong claim. It doesn't suffice, for instance, for p to be true in some columns and false in none. Each column has to take a stance on p, and endorse it.

But **Given** cannot be converted to a biconditional. Being disposed to sometimes take p as given is not sufficient for belief. If Anisa had played the Red-Green game rationally, she would have lost any belief about when

[1] Analysing these possibilities as sets of worlds is unhelpful when we want to use a model like this to represent modal or logical uncertainty. But it's a helpful heuristic in most cases. And there isn't anything wrong with using a model that breaks down when applied outside its appropriate zone.

the Battle of Hastings was. To explain cases like that, we need to expand our theory of belief.

3.4 Blocking Belief

Imagine a person, call him Erwan, who is made the offer Blaise is made, but declines it. He declines on the very sensible grounds that the Battle of Agincourt might not have been in 1415, and he does not want to run the risk of sending everyone to the Bad Place. If we stop our theory of belief with **Given**, then we have to say that Erwin has some kind of weird pragmatic incoherence. He believes that p, and wants what is best for everyone, but won't do the thing that will, given his beliefs, produce what is best for everyone. Declining the bet is not practically incoherent in this way. So Erwin does not believe that the Battle of Agincourt was in 1415. At least, he doesn't believe that at the time he is declining the bet.

So a theory of belief with any hope of being complete needs some supplementation. The idea I'll use is one that seems prima facie like it might apply without restriction. A little reflection, however, shows that it will ultimately need to be restricted, and the most natural restrictions are pragmatic.

Imagine that we don't ask Erwin whether he is prepared to bet the welfare of all of humanity on historical claims, but instead ask him a simple factual question (H).

(H) How many (full) centuries has it been since the Battle of Agincourt?

Erwin will think to himself, "Well, the Battle of Agincourt was in 1415, and that's a bit over 600 years ago, so that's six centuries. The answer is six." Now compare what happens if we ask him this slightly more convoluted question.

(I) If the Battle of Agincourt was in 1415, how many (full) centuries has it been since the Battle of Agincourt?

Erwin will give the same answer, i.e., six. And he will give it for basically the same reasons. Indeed, apart from the date of the Battle being one of his reasons in answering (H), and not needed to answer (I), he has the same reasons for answering the two questions with six. I mean that both in the sense that what justifies giving the answer six is the same for the two questions, and in the sense that what causes him to answer six is the same

for the two questions. (With the exception that the date of the Battle is a reason in answering (H), but not in answering (I).)

Say that a person answers the questions *Q?* and *If p, Q?* in the same way if they offer the same answer to the two questions, and their reasons (in both senses) for these answers are the same except only that *p* is one of the reasons for their answer to *Q?*. Then here is a plausible principle about belief - albeit one that isn't going to be quite right.

Unrestricted Conditional Questions If *S* believes that *p*, then for any question *Q?*, S is disposed to answer the questions *Q?* and *If p, Q?* the same way.

Note that in saying these questions are answered the same way, I really don't just mean that they get the same answers. I will offer the same answer to the questions *What is one plus one?* and *What is the largest n such that* $x^n + y^n = z^n$ *has positive integer solutions?*, but I don't answer these questions the same way. My reasons for the first answer are quite closely related to the fact that one plus one does equal two. My reasons for the second answer are almost wholly testimonial. So in the sense relevant to **Unconditional Conditional Questions**, I do not answer each question the same way.

I'm understanding what a conditional question is in a particular way. I think this is how conditional questions usually work in English, so the shorthand *If p, Q?* that I'm using is not misleading. But I don't intend to defend a particular claim about the way natural language conditionals work. That would be another whole book. (Or more.) So I intend to use this shorthand *If p, Q?* somewhat stipulatively.

If p, Q? is the question *Q?* asked under the assumption that *p* can be taken as given. So the question *If p, how probable is q?* is asking for the conditional probability of *q* given *p*. The question *If p, which option is most useful?* is asking for a comparison of the conditional utilities of the various options. And the question *If p, must it be that q?* gets an affirmative answer if all the (salient) possibilities where *p* is true are ones where *q* is true. (So it becomes very close to asking if the material implication $p \supset q$ must be true.) Now notoriously it is difficult to connect these conditional questions with questions about the truth of any conditional.[2] But I'm setting all those issues aside here. Everything that I say about conditional questions I could say, more verbosely, by making it explicit that they are

[2]See Lewis (1976; 1986) on the issues about conditional 'how probable' questions; Lewis (1988; 1996) on the issues about conditional 'how useful' questions; and Gillies (2010) on issues about modals in the consequent of conditional questions.

to be understood as questions about conditional probability, conditional utility, conditional modality, and so on.

Now thinking about a few simple cases might make it seem that **Unrestricted Conditional Questions** is true. After all, there is something very odd about a counterexample to it. It would have to be a case where *S* believes that *p*, and there is a way they are disposed to get answer *If p, Q?*, i.e., to get from *p* to an answer to *Q?*, but they are not disposed to use that to answer *Q?*. That seems at best rather odd.

Let me mention one potential counterexample that I don't think undermines **Unrestricted Conditional Questions**. There could be a case where I believe *p*, and *p* is relevant to *Q?*, but I don't realise its relevance. On the other hand, when I am explicitly asked *If p, Q?*, being reminded of *p* makes me see the connection, so I follow the natural path from *p* to an answer to *Q?*. These kind of one-off performance errors are, sadly, easy to make. But as long as they are one-off, they don't threaten the principle connecting dispositions.

But a bigger problem comes from the two cases that I started the book with. If the Battle of Agincourt was in 1415, then Anisa maximises expected utility by playing blue-true, and Blaise maximises expected utility by taking the bet. So answer to the conditional questions *If the Battle of Agincourt was in 1415, what options of Anisa's maximse expected utility?* and *If the Battle of Agincourt was in 1415, what option of Blaise's maximses expected utility?* have different answers to the corresponding unconditional questions. Or at least so say I, and hope you do too. So if **Unrestricted Conditional Questions** is true, then none of us have ever believed that the Battle of Agincourt was in 1415. That can't be right, so there must be some restriction on the principle.

Happily, a restriction isn't too hard to find. The principle just needs to be restricted to questions that the subject is currently taking an interest in. When we're thinking about questions like (H) and (I), then we do have beliefs about when the Battle of Agincourt was. Were we to be placed in Anisa or Blaise's situation, or arguably when we even think about their situation, we lose this belief. So I suggest the following principle is true, and explains a lot of the cases that have been discussed so far.

Relevant Conditional Questions If *S* believes that *p*, then for any question *Q?* that *S* is currently taking an interest in, *S* is disposed to answer the questions *Q?* and *If p, Q?* the same way.

As I argued in 2.5, whether one is interested in a question isn't just a matter of one's practical situation. One can be interested in a question because one is thinking about what to do should it arise, or because one is just naturally inquistive. Many of the questions we're interested in are practical questions, but not all of them are.

So I've argued that **Given** and **Relevant Conditional Questions** are necessary conditions on belief. And very roughly, I think they are jointly sufficient for belief. I say 'roughly' because I don't mean to take a stance on, say, whether animals have beliefs, or whether one can have singular thoughts about things one is not acquainted with. A more accurate claim is that if it is plausible that *S* is the kind of thing that can have beliefs, and *p* is the kind of thing it could in principle have beliefs about, and both **Given** and **Relevant Conditional Questions** are satisfied, then *S* believes that *p*.

Obviously neither **Given** and **Relevant Conditional Questions** would be particularly helpful principles to use in providing a reductive physicalist account of mental content. They say something about necessary conditions for belief, but the statement of those conditions makes a lot of assumptions about other content-bearing states of the agent. So even if these conditions are individually necessary and jointly sufficient for belief, they wouldn't be any kind of analysis or reduction of belief.[3] But they could be part of a theory of belief, and the theory they are part of is helpful for seeing how beliefs and interests fit together.

3.5 Questions and Conditional Questions

In the previous section I defended this principle:

Relevant Conditional Questions If *S* believes that *p*, then for any question *Q?* that *S* is currently taking an interest in, *S* is disposed to answer the questions *Q?* and *If p, Q?* the same way.

To spell out what that principle amounts to, I need to say something about what questions are, and what conditional questions are. I'm going to say just enough about questions to understand the principle. This won't be anything like a full theory of questions. While much of what I say will

[3]Compare: One can consistently deny that any analysis or reduction of *knowledge* is possible and say that the condition *p is part of S's evidence* is both necessary and sufficient for *S* to know that *p*.

draw on insights from theorists who have worked on questions in natural language, I'm not primarily interested in how questions are expressed in natural language. Rather, I'm interested in the contents of these questions. These contents are interesting because they can be the contents of mental states. For example, a cat can wonder where a mouse is hiding. There are deep and fascinating issues about how we can and do talk about the cat, and the cat's attitudes, but I'm more interested in the cat's relationship to the question *Where is the mouse hiding?* than I am in our talk about the cat.[4]

The simplest questions are true/false questions, like *Did the Boston Red Sox win the 2018 World Series?*. These won't play a huge role in what follows, but they are important to have on the table. I am going to assume that whenever someone considers a proposition, and they don't take its truth value to be settled, they are interested in the question of whether it is true.

Next, there are quantitative questions, where the answer is some number or sequence of numbers.[5] One tricky thing about quantitative questions is that they may admit of imprecise answers, but need not.

If I ask "When does tonight's Red Sox game start?", an answer of "Seven" would usually be acceptable, even if the game actually starts at a few minutes after seven. That's because, I take it, the truth conditional content of the utterance "Seven" in this context is that tonight's Red Sox game starts at approximately seven, and I'm asking a question that admits of an approximate answer. I could have been asking a question where the only acceptable answer would be the time that the Red Sox game starts to the nearest minute, or even to the nearest second. And I could even have asked that question using those exact same words. (Though if I intended to ask the question about seconds, using these words would be extremely unlikely to result in communicative success.)

The main thing that matters for the purposes of this book is that the questions with different appropriate answers are different questions. Even if one would normally use the same words in English to express the ques-

[4]A useful introduction to ways in which questions are relevant to philosophy of language is the Stanford Encyclopaedia article by Cross and Roelofsen (2018). A canonical text on the role of questions is Roberts (2012); this paper was originally circulated in 1996 and has influenced a huge range of works, including this one.

[5]I'm including here any question that could be answered with a number or sequence of numbers, even if that would not be the most usual, or the most helpful, way to answer them. So *Where is Fenway Park?* is a quantitative question, because *42.3467° N, 71.0972° W* is an answer, even if *The corner of Jersey St and Van Ness St* is a better answer.

tions, the fact that they have different acceptable answers shows that they are different questions.

And as noted above, what really matters for this book is the mental representation of the contents of questions. There could be two people who we could report as wondering when tonight's Red Sox game starts, but one of them will cease wondering if they find out that it starts around seven, and the other still wonders which minute near seven it will start at. These people are wondering about different questions.

The more precise a numerical question one is considering, the fewer things one can rationally take for granted in trying to answer it. So the version of IRT I defend implies that the more precise a numerical question one is considering, the fewer things one knows. Or, to put the same point another way, the less precise a numerical question one is considering, the less impact interest-relativity has on knowledge. This will matter when thinking about how the theory applies to various examples. If we have ascribe to a thinker an interest in an unrealistically precise question, we might draw implausible conclusions about what IRT says about them. But this isn't a consequence of IRT; it's a consequence of not getting clear about which question a thinker is considering.

Next, there are questions that ask to identify an individual or a class of individuals. A striking thing about these questions is that they often have so-called 'mention-some' readings. To understand what this means, compare these two little exchanges.

1. a. Who was in the Beatles?
   b. John Lennon was in the Beatles.

2. a. Where can I get good coffee in Melbourne?
   b. You can get good coffee at Market Lane.

There is something wrong with 1b as an answer to 1a. It's true that John Lennon was in the Beatles. But an ordinary use of 1a will be to ask for the names of everyone in the Beatles, not just one person in them. (There are exceptions, and it's a fascinating task to work out when they occur. But it's not my fascinating task.) On the other hand 2b is a perfectly good answer to 2a. (Or so I think, but my knowledge of Melbourne coffee is a little out of date.) It is definitely not necessary to properly answer 2a that one list every place in Melbourne where one can get good coffee. That could take some time. Moreover, 2b does not (on its most natural reading) imply that Market Lane is the only place in Melbourne to get good coffee.

An answer is a 'mention-some' answer when it does not imply exclusivity in this sense. And a question admits of mention-some answers when it is properly answered with a mention-some answer. Lots of questions asking for individuals will be mention-some questions in this sense, but not all of them will. And, again, it is important to understand what kind of question is being asked to think about whether it is satisfactorily answered by an answer that does not imply completeness or exhaustiveness.

Next, there are questions with infinitivals. In most dialects of English, it is rare to use these to simply ask questions.[6] But they can be the complements of any number of verbs. So here are some examples of what I mean by infinitival questions.

- When to visit Venice?
- How to climb Ben Nevis?
- What to do?

Any of these, and any number of other questions with infinitivals, can complete sentences like

- A doesn't know …
- B is wondering …
- C wants D to tell him …

Mixing and matching the sentence fragments from the last two lists produces nine different sentences. Some examples of these are

- C wants D to tell him how to climb Ben Nevis.
- A doesn't know what do do.
- B is wondering whether the visit Venice.

The philosophical work on these kinds of sentences has been almost exclusively focussed on just one of the nine sentences I just described: the one combining a knowledge verb with a 'how to' question. I suspect this is a mistake; what to say about 'know how' reports is going to have a lot in common with what to say about 'wondering when' reports. (Here I'm agreeing with Stanley (2011), though I'm about to disagree with him on a related point.)

There is a puzzle about why, in English, we cannot use these questions to complete sentences like

[6] My hunch is that there is quite a bit of dialectical variation here, I would need to do much more empirical research to back this up.

- E believes …
- F suspects …
- G wants H to guess …

I'm going to set that puzzle aside, as interesting as it is, an just focus on the sentences we can produce in English.

I'm going to call these questions with infinitivals *practical questions*. One thing to note aboutis that they are are usually mention-some. When I am wondering what to buy, and I resolve this by choosing one particular carton of eggs, I don't thereby imply that there is anything defective about the other cartons. I just choose some eggs.

For related reasons, answering a practical question like this is distinct from answering any question, or questions, about the modal status of different actions. Imagine that in the grip of choice-phobia I am stuck staring at the cartons of eggs, unable to decide which one to buy because they are all just alike. In that situation I might know that there is no carton such that it is what I should buy, and also that there are many cartons such that I could (rationally, morally) buy any one of them. But there are so many, and they are so alike and I can't decide, so I don't know what to buy.

Resolving this indecision will not involve accepting any modal proposition like *I should buy this carton in particular*. It better not, because I really have no reason to accept any such proposition. Rather, it involves accepting a proposition like *I will buy this carton in particular*. And that I can accept by simply buying the eggs. But there were many answers I could equally well have accepted, since there were many other cartons I could buy.[7]

Practical questions are distinct from questions about modals or utilities. But there will usually be a correlation between their answers. Usually, if someone asks you when to visit Venice, and there is one time in particular such that visiting then maximises expected utility, that's what you should tell them. That's when they should visit, and that's what to say when they ask you when to vist. Relatedly, practical questions can come in conditional form. We can utter sentences like the following in English.

- J asks K what to do if his patient has hepatitis.

[7] I'm here mildly disagreeing with Jason Stanley (2011, Ch. 5) when he says that these questions with infinitival complements can be paraphrased using modals. But only mildly since (a) we might think 'will' just is the modal that gets used in the paraphrase, as Bhatt (1999) suggests, and (b) the differences between *what to buy* and *what may I buy* are small enough that maybe we can describe the latter as a 'paraphrase' of the former.

And there is one feature of these sentences that needs noting. I don't know what to do if one's patient has hepatitis, so let's just say that J tells K to do X. What that means is not that in any situation where the patient has hepatitis, do X. If the patient's symptoms are confusing, it might be best to run more tests before doing X. What it does mean is that if the fact that the patient has hepatitis is taken as given, then do X. As always, conditional questions should be understood as questions about what happens in scenarios where the condition in question is taken as given. And the constraint expressed by **Relevant Conditional Questions** is that whatever is known can be taken as given in just this sense.

3.6 A Million Dead End Streets

As I've noted already, the view I'm defended here is somewhat different from my earlier view. And it's helpful too understand the view of this book to lay out, in one place, the ways in which time has changed my views. Here is a somewhat simplified version of the view from *Can We Do Without Pragmatic Encroachment*. Assume that S is interested in some quantitative questions and some alethic (i.e., yes/no) questions. Then the view was that S believes that p if and only if these two conditions are met.

1. For any quantitative question *Q?* that S is interested in, and any alethic question A that S is interested in, S's answers to the question *If A, Q?* and *If A and p, Q?* are the same.
2. S's credence in p is greater than 0.5.

It was assumed that S is always 'interested' in the null question *Is a tautology true?*, so one special instance of this is that S answers *Q?* and *If p, Q?* the same way. And it was assumed that S is an expected utility maximiser, so the practical question of what to do becomes just the quantitative question *Which of these options has the highest expected utility?*. There are bells and whistles, especially in thinking about the level of precision that goes along with the quantitative questions that S is interested in. (Draw these too fine, and S doesn't have beliefs, so you have to be a little careful here.) But this is enough to see the basic view, and to see its problems.

There is a lot about this view I've kept. There are still the two parts, one primarily to do with propositions that are not practically relevant, and one to do with propositions that are. And the clause to deal with propositions that are practically relevant requires close match between conditional and unconditional answers.

But there are a lot of changes. Going from saying that the credence in p is greater than 0.5 to saying that S is willing, at least sometimes, to take p for granted, is a big change. I no longer presuppose that questions about what to do just are questions about expected utility. I've stopped focussing exclusively on answers to (conditional) questions, and moved to talking about both answers and ways that questions are answered. And I dropped the requirement that we look at these potentially quite abtruse questions, such as how to answer $Q?$ assuming both A and p. The last two changes offset each other; the reason for including these doubly conditional questions was, in effect, to look at how S was willing to get to answers about questions with more practical import.

There are many reasons, most of them due to perceptive critics of my earlier work, for making these changes. I'll just focus here on the five that have been most significant.

3.6.1 Correctness

Ross and Schroeder (2014) note that my earlier theory doesn't have a good story about why false beliefs are incorrect.[8] I think that's right. Even if p is false, there is nothing necessarily mistaken about either having credence in p above 0.5, or in having unconditional preferences match preferences conditional on p.

But surely false beliefs are, in a way, incorrect. They may be rational, they may be well-supported, and so on, but still if you believe that p, and p turns out not to be the case, you got it wrong. There are other mental states that have truth as a correctness condition. Guesses are correct or incorrect, even if there need be nothing at all irrational about making a false guess. Indeed, any mortal who doesn't make false guesses from time to time isn't playing the guessing game well. But not all mental states are like this. If I hope that p, and p doesn't come to pass, that doesn't make my hope incorrect. It just makes it frustrated. So to say that a false belief is incorrect is not to just make the trivial point that it is false. It is also to say that the belief failed to meet one important standard of evaluation for beliefs - correctly representing the world.

The new theory does not have this problem. Doing dominance reasoning where all of the situations one considers are non-actual is a mistake. It's not a mistake because it will inevitably lead to an irrational decision.

[8] Fantl and McGrath (2009) make a similar argument, targetted at Lockean theories of belief more than at my theory. I'll come back to how this is a problem for Lockean theories in subsection 6.4.2.

But it's a mistake because one draws a conclusion that is not supported by the premises it is based on. Those premises only say that one option is better than another conditional on one or other condition obtaining. And that's a bad reason to say the first option is simply better if there is some extra option that might obtain. And whatever does obtain, might obtain.

This way of explaining the incorrectness of false belief suggests a central role for knowledge in norms of beliefs. False beliefs are mistaken because they lead one to treat the actual situation as one that could not obtain, yet the actual situation might obtain. But one can make the same mistake by treating a situation that doesn't obtain, but might, as one that could not obtain. And believing something one doesn't know will (typically) lead to doing that.

3.6.2 Impractical Propositions

The second clause in my earlier theory was designed to rule out trivial belief in irrelevant propositions. The first clause on its own has some absurd consequences. Imagine that I'm relaxing by a stream watching the ripples without a care in the world. All of the very few questions that I'm currently interested in have the same answer unconditionally as they do conditional on the Battle of Agincourt having been fought in 1415. So according to clause 1, I believe the Battle of Agincourt was in 1415. That's good, because I do believe that. But it's also true that all of the very few questions that I'm currently interested in have the same answer unconditionally as they do conditional on the Battle of Agincourt having been fought in 1416. So if clause 1 was the full theory of belief, then I would also believe that the Battle of Agincourt was in 1416. And that's false.

So I added clause 2 to the theory in order to fix this problem. But it only fixes a special case of the problem. Let p be the proposition that the next die I roll will land 1, 2, 3 or 4. My credence in that is $\frac{2}{3}$, so it satisfies clause 2. And conditionalising on it doesn't change the answer to any of the very few problems that I'm interested while the ripples float down the stream. So I believe p. But that's absurd too. (This objection is also due in important parts to Ross and Schroeder (2014), though my presentation differs from theirs to emphasise just which parts of the objections most worry me.)

The new theory handles this case easily. There is no context where I would simply ignore the possibility that this next die roll will land 5 or 6 for the purposes of doing dominance reasoning. So I don't believe that p, as required.

Is there anything we can rule out on purely probabilistic grounds? It's a little interesting to think this kind of case through. Imagine there is some salient very large number, and it matters what the remainder is when that large number is divided by 1000, or 1000. Could we get to a point where a choice that is better than some alternative unless that remainder is, say 537, feel like a dominating choice? I'm not sure whether that would ever happen. But it does seem plausible to say that whether such a choice ever feels like a dominating choice correlates with whether we could ever straight up believe that the remainder is not precisely 537 on purely probabilistic grounds.

3.6.3 Choices with More Than Two Options

Consider this variant of the Red-Blue game. As well as the four options Anisa has in the original version of the game, she has a fifth option. This option says that if she answers some question correctly, she wins $100. She's told what the question is, and what the red and blue sentences are, before she has to choose. And in this case, the question is, who was the first American woman to win an Olympic gold medal.

Imagine that Anisa just skim reads the red and blue sentences, and doesn't think about which of them she'd pick, because she knows the answer to this question. It was, she knows, Margaret Abbott. So she promptly gives that answer, and wins $100.

Now she clearly takes an interest in the options Red-True and Blue-True. She has reasons for preferring to answer the question than take one of those two options. And she could give those reasons without any reflection. So Red-True and Blue-True should be in the range of things that we quantify over when thinking about options she is interested in. And she has a stable disposition to choose Red-True over Blue-True; I think that stable disposition is a strict preference. And that strict preference does not survive conditionalising on the proposition that the Battle of Avignon was in 1415. So my earlier theory says that even in this revised version of the game, Anisa does not believe that the Battle of Avignon was in 1415.

And this seems mistaken to me. In any deliberation Anisa does, her regular disposition to take it for granted that the Battle of Avingon was in 1415 survives. There is a very nearby deliberation where it does not survive, namely the deliberation about whether Red-True or Blue-True is better. But, crucially, she does not have to take an interest in that question in order to take an interest in the two options Red-True and Blue-True. If they are both (clearly) suboptimal options in her current situation, she

can simply settle for concluding that they are suboptimal, and leave it at that.

So I think my old theory made it too easy to lose belief in cases where one has to choose between many options. Being interested in some options, because you want to choose the best one of them, does not mean being interested in all questions about preferences between pairs of them. The problem was that I'd been focussing largely on two-way choices, so the distinction between being interested in some choices and being interested in which of those two is better got elided. But that distinction matters, and the hybrid pragmatic theory handles it better than my old theory.

3.6.4 Hard Times and Close Calls

In my earlier theory, any practical deliberation was modelled as an inquiry into which option had the highest expected utility. This was wrong for a number of reasons, not least that it gives implausible results in cases involving choices between very similar options. I'll briefly describe one example that illustrates the problem, and the start of how I plan to solve it. But it turns out to be rather tricky to get the details right, and I'll come back to this in subsection 4.3.1 and again in chapter 7. The details of the example are new, but it's a very minor modification of a kind of example that is discussed in McGrath and Kim (2019) and credited to a talk by John Hawthorne "circa 2007". Similar examples are also discussed by Alex Zweber (2016) and by Anderson and Hawthorne (2019b), and I'm drawing on their insights in describing this one.

David is doing the weekly groceries. He needs a can of chickpeas, so he walks to where the chickpeas are and looks at the shelf. There are two cans, call them c_1 and c_2, that are equally easy to reach and get from the shelf. Call the actions of taking them t_1 and t_2. David simply assumes, partially on inductive grounds and partially on grounds of what he knows about supermarkets, that neither can has passed its expiry date. But while it is wildly implausible that either can has, the probability is not zero. Let e_i be that can i has expired, and assume that $\Pr(e_1)$ and $\Pr(e_2)$ are low and equal. Call this probability e. Let h be the utility of choosing an unexpired can, and l the utility of choosing an expired can, where obviously $h > l$. Then both t_1 and t_2 have utility $(1-e)h+el$. Conditional on $\neg e_1$, the utility of t_1 is h, which is greater than $(1-e)h + el$ as long as $e > 0$ and $h > l$. So unconditionally, t_1 and t_2 have the same utility, but conditional on $\neg e_1$, they have different utilities. So, according to the theory I used to defend, when David is making this choice, he does not believe, and hence does

not know $\neg e_1$. This seems wrong, and there are even worse consequences one can draw my thinking about minor variants of the case.

The key part of my response to this will be distinguishing between the questions *Which can to choose?*, and the question *Which choice of can has maximal expected utility?*. If David is thinking about the latter question, then it turns out he really doesn't know $\neg e_1$. That's a somewhat surprising result, and I'll turn to defending it in chapter 7. But as long as he is focussing solely on the former question, the argument of the previous paragraph doesn't go through.

So the big move here is to move from somewhat quantitative questions, like *Which choice maximises expected utility?*, to practical questions like *What to do?*. Once we do that, the problem that Zweber, and Anderson and Hawthorne, raise ceases to be a problem. I don't intend these brief remarks to be a convincing case that I've got a good solution to these problems. Rather, the point is to flag that the theory I'm defending here is distinct from the theory I used to defend, and this gives me some more resources to handle cases like David and the chickpeas.

3.6.5 Updates and Modals

The version of IRT that I defend here gives a big role to conditional attitudes.[9] That's something that it has in common with everything I've written about IRT. But I used to have a particular pair of views about how to understand conditional attitudes. In particular, I took the following two claims to be at least close approximations to the truth about conditional attitudes.

1. An attitude conditional on p is (usually) the same as the attitude one would have after updating on p.
2. The way to update on p is to conditionalise.

The first is at best an approximation for familiar reasons. I can think that no one knows whether p is true, and even think that this is true conditional on p. But after updating on p, I will no longer think that. So we have to be a bit careful in applying principle 1; it has counterexamples. But it is still a useful heuristic, and that's how I'll use it.

What wasn't originally obvious to me was that there are counterexamples to principle 2 as well. And they are more significant for the way IRT

[9] This section is based on material from my (2016, sect. 1).

should be understood. I used to describe the picture of belief I was defending as the view that to believe something is to have a credence in it that's close enough to 1 for current purposes. That's still a decent heuristic, but it isn't always right. When someone is interested in modal questions, credence 1 might be insufficient for belief. To see how this might be so, it helps to start with some points Thony Gillies (2010) makes about the relationship between modals, conditionals and updating.

When modal questions are on the table, updating will not be the same as conditionalising. This is shown by the following example. (A similar example is in Kratzer (2012, 94).)

> I have lost my marbles. I know that just one of them – Red or Yellow – is in the box. But I don't know which. I find myself saying things like …"If Yellow isn't in the box, the Red must be." (4:13)

What matters for the purposes of this book is not whether this conditional is true, but whether its truth is consistent with the Ramsey test view of conditionals. And Gillies argues that it is.

> The Ramsey test – the schoolyard version, anyway – is a test for when an indicative conditional is acceptable given your beliefs. It says that (if p)(q) is acceptable in belief state B iff q is acceptable in the derived or subordinate state B-plus-the-information-that-p. (4:27)

And he notes that this can explain what goes on with the marbles conditional. Add the information that Yellow isn't in the box, and it isn't just true, but must be true, that Red is in the box.

Note though that while we can explain this conditional using the Ramsey test, we can't explain it using any version of the idea that probabilities of conditionals are conditional probabilities. The probability that Red must be in the box is 0. The probability that Yellow isn't in the box is not 0. So conditional on Yellow not being in the box, the probability that Red must be in the box is still 0. Yet the conditional is perfectly assertable.

There is, and this is Gillies's key point, something about the behaviour of modals in the consequents of conditionals that we can't capture using conditional probabilities, or indeed many other standard tools. And what goes for consequents of conditionals goes for updated beliefs too. Learn that Yellow isn't in the box, and you'll conclude that Red must be. But

that learning can't go via conditionalisation; just conditionalise on the new information and the probability that Red must be in the box goes from 0 to 0.

Now it's a hard problem to say exactly how this alternative to updating by conditionalisation should work. But very roughly, the idea is that at least some of the time, we update by eliminating worlds from the space of possibilities. This affects dramatically the probability of propositions whose truth is sensitive to which worlds are in the space of possibiilties.

And this matters when we are considering modal questions. For example, if we are considering the question *Must q be true?*, then it is plausible that unconditionally the answer is no, and indeed the unconditional probability that q must be true is 0, but that conditional on p, q must be true.

We don't even have to be considering modals directly for this to happen. Assume that actions A and B have the same outcome conditional on q, but A is better than B in every $\neg q$ possibility. Then if we are considering the question *Is A better than B?*, it will matter whether it must be the case that q.

Assume that q could have probability 1 without it being the case that q must be true. (This is controversial, but I'll offer arguments in sections ?? and 6.3 that it is possible.) Then unconditionally, A is better than B, even though they have the same expected utility. That's because weak dominance is a good principle of practical reasoning: If A might be better than B and must not be worse, then A is better than B. But by hypothesis, conditional on p, A is not better than B. So in this case p will not be believed; conditional on p the question *Is A better than B* gets a different answer to what it gets unconditionally.

Note though that all I said to get this example going is that p rules out $\neg q$, and q has probability 1. That means p could have any probability at all, up to probability 1. So it's possible that conditional on p, some relevant questions get different answers to what they get unconditionally, even though p has probability 1. So belief can't be a matter of having probability close enough to 1 for practical purposes; sometimes even probability 1 is insufficient.

3.7 Nearby Views

3.7.1 Ganson's Theory

3.7.2 Ross and Schroeder's Theory

Jacob Ross and Mark Schroeder (2014) have what looks like, on the surface, a rather different view to mine.[10] They say that to believe p is to have a **default reasoning disposition** to use p in reasoning. Here's how they describe their view.

> What we should expect, therefore, is that for some propositions we would have a *defeasible* or *default* disposition to treat them as true in our reasoning–a disposition that can be overridden under circumstances where the cost of mistakenly acting as if these propositions are true is particularly salient. And this expectation is confirmed by our experience. We do indeed seem to treat some uncertain propositions as true in our reasoning; we do indeed seem to treat them as true automatically, without first weighing the costs and benefits of so treating them; and yet in contexts such as High where the costs of mistakenly treating them as true is salient, our natural tendency to treat these propositions as true often seems to be overridden, and instead we treat them as merely probable.
>
> But if we concede that we have such defeasible dispositions to treat particular propositions as true in our reasoning, then a hypothesis naturally arises, namely, that beliefs consist in or involve such dispositions. More precisely, at least part of the functional role of belief is that believing that p defeasibly disposes the believer to treat p as true in her reasoning. Let us call this hypothesis the *reasoning disposition account* of belief. (Ross and Schroeder, 2014, 9-10)

There are, relative to what I'm interested in, three striking characteristics of Ross and Schroeder's view.

1. Whether you believe p is sensitive to how you reason; that is, your theoretical interests matter.
2. How you would reason about some questions that are not live is relevant to whether you believe p.

[10] This is based on material in my (2016, sect. 3).

3. Dispositions can be masked, so you can believe p even though you don't actually use p in reasoning now.

The view I'm defending here agrees with them about 1 and 2, though my theory manifests those characteristics in a quite different way. But point 3 is a cost of their theory, not a benefit, so it's good that my theory doesn't accommodate it. (For the record, the theory I put forward in my (2005a) did not agree with them on point 2, and I changed my view because of their arguments.)

I agree with 1 because, as I've noted a few times above, I think theoretical interests as well as pragmatic interests matter for the relationship between credence and belief. And I agree with 2 because I think that whether someone is disposed to use p as a premise matters to whether they believe p. Let p be some ordinary proposition about the world that a person believes, such as that the Florida Marlins won the 2003 World Series. And let q be a lottery proposition that is just as probable as p. (That is, let q be a lottery proposition such that if the person were to play the Red-Blue game with p as red and q as blue, they would be rationally indifferent between the choices.) Then on my theory the person believes p but not q, and this isn't due to any features of their credal states. Rather, it is due to their dispositions to use p as a premise in reasoning. (For example, they might use it in figuring out how many World Series were won by National League teams in the 2000s.)

Ross and Schroeder argue, and I basically agree, that interest-relative theories of belief that only focus on practical interests have trouble with folks who use odd techniques in reasoning. This is the lesson of their example of *Renzi*. I'll run through a somewhat more abstract version of that case, because the details are not particularly important. Start with a standard decision problem. The agent knows that X is better to do if p, and Y is better to do if $\neg p$. The agent should then go through calculating the relative gains to doing X or Y in the situations they are better, and the probability of p. But the agent imagined doesn't do that. Rather, the agent divides the possibility space in four, taking the salient possibilities to be $p \wedge q$, $p \wedge \neg q$, $\neg p \wedge q$ and $\neg p \wedge \neg q$, and then calculates the expected utility of X and Y accordingly. This is a bad bit of reasoning on the agent's part. In the cases we are interested in, q is exceedingly likely. Moreover, the expected utility of each act doesn't change a lot depending on q's truth value. So it is fairly obvious that we'll end up making the same decision whether we take the 'small worlds' in our decision model to be just the world where p, and the world where $\neg p$, or the four worlds this agent uses.

But the agent does use these four, and the question is what to say about them.

Ross and Schroeder say that such an agent should not be counted as believing that q. If they are consciously calculating the probability that q, and taking $\neg q$ possibilities into account when calculating expected utilities, they regard q as an open question. And regarding q as open in this way is incompatible with believing it.

I agree. The agent was trying to work out the expected utility of X and Y by working out the utility of each action in each of four 'small worlds', then working out the probability of each of these. Conditional on q, the probability of two of them ($p \wedge \neg q$, $\neg p \wedge \neg q$), will be 0. Unconditionally, this probability won't be 0. So the agent has a different view on some question they have taken an interest in unconditionally to their view conditional on q. So they don't believe q. The agent shouldn't care about that question, and conditional on each question they should care about, they have the same attitude unconditionally and conditional on q. But they do care about these probabilistic questions, so they don't believe q. (And again for the record, the theory I defended at the time Ross and Schroeder wrote their paper did not have the resources to make this reply; I've changed my views in light of their arguments.)

So far I've been agreeing with Ross and Schroeder. But there is one big point of disagreement. They think it is very important that a theory of belief vindicate a principle they call **Stability**.

> **Stability**: A fully rational agent does not change her beliefs purely in virtue of an evidentially irrelevant change in her credences or preferences. (2014, 20)

Here's the kind of case that is meant to motivate Stability, and show that views like mine are in tension with it.

> Suppose Stella is extremely confident that steel is stronger than Styrofoam, but she's not so confident that she'd bet her life on this proposition for the prospect of winning a penny. PCR [their name for my old view] implies, implausibly, that if Stella were offered such a bet, she'd cease to believe that steel is stronger than Styrofoam, since her credence would cease to rationalize acting as if this proposition is true. (2014, 20)

Ross and Schroeder's own view is that if Stella has a defeasible disposition to treat as true the proposition that steel is stronger than Styrofoam, that's

enough for her to believe it. And that can be true if the disposition is not only defeasible, but actually defeated in the circumstances Stella is in. This all strikes me as just as implausible as the failure of Stability. Let's go over its costs.

The following propositions are clearly not mutually consistent, so one of them must be given up. We're assuming that Stella is facing, and knows she is facing, a bet that pays a penny if steel is stronger than Styrofoam, and costs her life if steel is not stronger than Styrofoam.

1. Stella believes that steel is stronger than Styrofoam.
2. Stella believes that if steel is stronger than Styrofoam, she'll win a penny and lose nothing by taking the bet.
3. If 1 and 2 are true, and Stella considers the question of whether she'll win a penny and lose nothing by taking the bet, she'll believe that she'll win a penny and lose nothing by taking the bet.
4. Stella prefers winning a penny and losing nothing to getting nothing.
5. If Stella believes that she'll win a penny and lose nothing by taking the bet, and prefers winning a penny and losing nothing to getting nothing, she'll take the bet.
6. Stella won't take the bet.

It's part of the setup of the problem that 2 and 4 are true. And it's common ground that 6 is true, at least assuming that Stella is rational. So we're left with 1, 3 and 5 as the possible candidates for falsehood.

Ross and Schroeder say that it's implausible to reject 1. After all, Stella believed it a few minutes ago, and hasn't received any evidence to the contrary. And I guess rejecting 1 isn't the most intuitive philosophical conclusion I've ever drawn. But compare the alternatives!

If we reject 3, we must say that Stella will simply refuse to infer r from p, q and $(p \wedge q) \rightarrow r$. Now it is notoriously hard to come up with a general principle for closure of beliefs. But it is hard to see why this particular instance would fail. And in any case, it's hard to see why Stella wouldn't have a general, defeasible, disposition to conclude r in this case, so by Ross and Schroeder's own lights, it seems 3 should be acceptable.

That leaves 5. It seems on Ross and Schroeder's view, Stella simply must violate a very basic principle of means-end reasoning. She desires something, she believes that taking the bet will get that thing, and come with no added costs. Yet, she refuses to take the bet. And she's rational to do so! At this stage, I think I've lost what's meant to be belief-like about

their notion of belief. I certainly think attributing this kind of practical incoherence to Stella is much less plausible than attributing a failure of Stability to her.

Put another way, I don't think presenting Stability on its own as a desideratum of a theory is exactly playing fair. The salient question isn't whether we should accept or reject Stability. The salient question is whether giving up Stability is a fair price to pay for saving basic tenets of means-end rationality. And I think that it is. Perhaps there will be some way of understanding cases like Stella's so that we don't have to choose between theories of belief that violate Stability constraints, and theories of belief that violate coherence constraints. But I don't see one on offer, and I'm not sure what such a theory could look like.

I have one more argument against Stability, but it does rest on somewhat contentious premises. There's often a difference between the best methodology in an area, and the correct epistemology of that area. When that happens, it's possible that there is a good methodological rule saying that if such-and-such happens, re-open a certain inquiry. But that rule need not be epistemologically significant. That is, it need not be the case that the happening of such-and-such provides evidence against the conclusion of the inquiry. It just provides a reason that a good researcher will re-open the inquiry. And, as I've argued above, an open inquiry is incompatible with belief.

Here's one way that might happen. Like other non-conciliationists about disagreement, e.g., Kelly (2010), I hold that disagreement by peers with the same evidence as you doesn't provide *evidence* that you are wrong. But it might provide an excellent reason to re-open an inquiry. We shouldn't draw conclusions about the methodological significance of disagreement from the epistemology of disagreement. So learning that your peers all disagree with a conclusion might be a reason to re-open inquiry into that conclusion, and hence lose belief in the conclusion, without providing evidence that the conclusion is false. This example rests on a very contentious claim about the epistemology of disagreement. But any gap that opens up between methodology and epistemology will allow such an example to be constructed, and hence provide an independent reason to reject Stability.

3.7.3 Leitgeb's Stability Theory

3.8 Weak Belief

Chapter 4

Knowledge

In chapter 3, I argued that to believe something is to take it as given in all relevant inquiries, and in at least one possible inquiry. And I explained what it was to take something as given in terms of how one answers conditional and unconditional questions. In this chapter I'm going to argue that whatever is known can be properly taken as given in all relevant inquiries. Since some things that are usually known cannot be properly taken as given in some inquiries, this implies that knowledge is relevant to one's inquiries and hence to one's interests.

There is an easy argument for the conclusion of this chapter.

1. To believe something is to, inter alia, take it as given for all relevant inquiries.
2. Whatever is known is correctly believed.
3. So, whatever is known is correctly taken as given in all relevant inquiries.

I think this argument is basically sound. But both premises are controversial, and it isn't completely obvious that it is even valid. So I'm not going to rely on this argument. Rather, I'll argue more directly for the conclusion that whatever is known is correctly taken as given in all relevant inquiries. This will provide indirect evidence that the theory of belief in chapter 3 was correct, since we can now take that theory of belief to be an explanation for the claim that whatever is known is correctly taken as given in all relevant inquiries, rather than as part of the motivation for it.

The argument here will be in two parts. First, I'll focus on practical inquiries, i.e., inquiries about what to do, and argue that what is known can be taken as given in all practical inquiries. Then I'll extend the discussion

to theoretical inquiries, and hence to inquiries in general. Then with the argument complete, I'll look at two possible objections to the argument - that it has implausible consequences about the role of logical reasoning in extending knowledge, and that it leads to implausible results when a source provides both relevant and irrelevant information.

4.1 Knowledge and Practical Interests

A practical inquiry can usually be represented by the kind of decision table that we use in decision theory courses.[1] We take something as given iff it is encoded in the right way in the decision table. The primary way that we encode a proposition p into a decision table is to set up the table so that p is true in every column. If we use a table where p is encoded in this way, and p is not known, we are making a mistake. And in particular, we are making an epistemic mistake.

To see this, let's start with an example. Professor Dec is teaching introductory decision theory to her undergraduate class. She is trying to introduce the notion of a dominant choice. So she introduces the following problem, with two states, S_1 and S_2, and two choices, C_1 and C_2, as is normal for introductory problems.

	S_1	S_2
C_1	-$200	$1000
C_2	-$100	$1500

She's hoping that the students will see that C_1 and C_2 are bets, but C_2 is clearly the better bet. If S_1 is actual, then both bets lose, but C_2 loses less money. If S_2 is actual, then both bets win, but C_2 wins more. So C_2 is better. That analysis is clearly wrong if the state is causally dependent on the choice, and controversial if the states are evidentially dependent on the choices. But Professor Dec has not given any reason for the students to think that the states are dependent on the choices in either way, and in fact the students don't worry about that kind of dependence.

That doesn't mean, however, that the students all adopt the analysis that Professor Dec wants them to. One student, Stu, is particularly unwilling to accept that C_2 is better than C_1. He thinks, on the basis of his experience, that when more than $1000 is on the line, people aren't as reliable about paying out on bets. So while C_1 is guaranteed to deliver $1000

[1] This section is based on my (2012, sect 1.1).

if S_2, if the agent bets on C_2, she might face some difficulty in collecting on her money.

Given the context, i.e., that they are in an undergraduate decision theory class, it seems that Stu has misunderstood the question that Professor Dec intended to ask. But it is not easy to specify just exactly what Stu's mistake is. It isn't that he thinks Professor Dec has misdescribed the situation. It isn't that he thinks the agent won't collect $1500 if she chooses C_2 and is in S_2. He just thinks that she *might* not be able to collect it, so the expected payout might really be a little less than $1500.

But Stu is not the only problem that Professor Dec has. She also has trouble convincing Dom of the argument. He thinks there should be a third state added to the table, S_3. In S_3, there is a vengeful God who is about to end the world, and take everyone who chose C_1 to heaven, while sending everyone who chose C_2 to hell. Since heaven is better than hell, C_2 does not dominate C_1; it is worse in S_3. Dom does not think this is particularly likely, but he thinks it is possible, and decision theory should represent possibilities like this. If decision theory is to be useful, we must say something about why we can leave states like S_3 off the decision table.

So in order to teach decision theory, Professor Dec has to answer two questions.

1. What makes it legitimate to write something on the decision table, such as the '$1500' we write in the bottom right cell of Dec's table?
2. What makes it legitimate to leave something off a decision table, such as leaving Dom's state S_3 off the table?

When we've talked about decision tables so far, the focus has been on what tables thinkers actually use. Here we are switching the focus to talk about what tables they should use. And the claim is going to be that what they should use is determined by what they know.

To get to that conclusion, start with a much simpler problem. Mireille is out of town on a holiday, and she faces the following decision choice concerning what to do with a token in her hand.

Choice	**Outcome**
Put token on table	Win $1000
Put token in pocket	Win nothing

This looks easy, especially if we've taken Professor Dec's class. Putting the token on the table dominates putting the token in her pocket. It returns

$1000, versus no gain. So she should put the token on the table.

I've left Mireille's story fairly schematic; let's fill in some of the details. Mireille is on holiday at a casino. It's a fair casino in the sense that the probabilities of the outcomes of each of the games is just what you'd expect. And Mireille knows this. The table she's standing at is a roulette table. The token is a chip from the casino worth $1000.

Putting the token on the table means placing a bet. As it turns out, it means placing a bet on the roulette wheel landing on 28. If that bet wins she gets her token back and another token of the same value. There are many other bets she could make, but Mireille has decided not to make all but one of them. Since her birthday is the 28th, she is tempted to put a bet on 28; that's the only bet she is considering. If she makes this bet, the objective chance of her winning is 1/38, and she knows this. As a matter of fact she will win, but she doesn't know this. (This is why the description in the table I presented above is truthful, though frightfully misleading.) As you can see, the odds on this bet are terrible. She should have a chance of winning around 1/2 to justify placing this bet. (It's a very unfair casino in this sense, but what can you expect at a vacation resort?) So the above table, which makes it look like placing the bet is the dominant, and hence rational, option, is misleading.

Just how is the table misleading though? It isn't because what is says is false. If Mireille puts the token on the table she wins $1000; and if she doesn't, she stays where she is. It isn't, or isn't just, that Mireille doesn't believe the table reflects what will happen if she places the bet. As it turns out, Mireille is smart, so she doesn't form beliefs about chance events like roulette wheels. But even if she did, that wouldn't change how misleading the table is. The table suggests that it is rational for Mireille to put the token on the table. In fact, that is irrational. And it would still be irrational if Mireille believes, irrationally, that the wheel will land on 28.

A better suggestion is that the table is misleading because Mireille doesn't know that it accurately depicts the choice she faced. If she did know that these were the outcomes to putting the token on the table versus in her pocket, it would be rational for her to put it on the table. If we take it as understood in a presentation of a decision problem that the agent knows that the table accurately depicts the outcomes of various choices in different states, then we can tell a plausible story about the miscommunication between Professor Dec and her students. Stu was assuming that if the agent wins $1500, she might not be able to easily collect. That is, he was assuming that the agent does not know that she'll get $1500 if she

chooses C_2 and is in state S_2. Professor Dec, if she's anything like other decision theory professors, will have assumed that the agent did know exactly that. And the miscommunication between Professor Dec and Dom also concerns knowledge. When Dec wrote that table up, she was saying that the agent knew that S_1 or S_2 obtained. And when she says it is best to take dominating options, she means that it is best to take options that one knows to have better outcomes. So here are the answers to Stu and Dom's challenges.

- It is legitimate to write something on the decision table, such as the '$1500' we write in the bottom right cell of Dec's table, iff the decision maker knows it to be true.
- It is legitimate to leave something off a decision table, such as leaving Dom's state S_3 off the table, iff the decision maker knows it not to obtain.

Perhaps those answers are not correct, but what we can clearly see by reflecting on these cases is that the standard presentation of a decision problem presupposes not just that the table states what will happen, but the agent stands in some special doxastic relationship to the information explicitly on the table (such as that the chooser in Professor Dec's example will get $1500 if C_2 and S_2) and implied by where the table ends (such as that S_3 will not happen).

I think that special doxastic relationship is knowledge. But I don't need to argue for that here. All I need to argue is that if the person making the decision knows that p, she stands in the special relationship.

But could the 'special doxastic relationship' be stronger than knowledge? Could it be, for example, that the relationship is certainty, or some kind of iterated knowledge? Plausibly in some game-theoretic settings it is stronger - it involves not just knowing that the table is accurate, but knowing that the other player knows the table is accurate. In some cases, the standard treatment of games will require positing even more iterations of knowledge. For convenience, it is sometimes explicitly stated that iterations continue indefinitely, so each party knows the table is correct, and knows each party knows this, and knows each party knows that, and knows each party knows *that*, and so on. An early example of this in philosophy is in the work by David Lewis (1969) on convention. But it is usually acknowledged (again in a tradition extending back at least to Lewis) that only the first few iterations are actually needed in any problem, and it seems a mistake to attribute more iterations than are actually

used in deriving solutions to any particular game.

The reason that would be a mistake is that we want game theory, and decision theory, to be applicable to real-life situations. There is very little that we know, and know that we know, and know we know we know, and so on indefinitely (Williamson, 2000, Ch. 4). There is, perhaps, even less that we are certain of. If we only could say that a person is making a particular decision when they stand in these very strong relationships to the parameters of the decision table, then people will almost never be making the kinds of decision we study in decision theory. Since decision theory and game theory are not meant to be that impractical, the 'special doxastic relationship' cannot be that strong. It could be that in some games, the special relationship will involve a few iterations of knowledge, but in decision problems, where the epistemic states of others are irrelevant, even that is unnecessary, and simple knowledge seems sufficient.

It might be argued here that we shouldn't expect to apply decision theory directly to real-life problems, but only to idealised versions of them, so it would be acceptable to, for instance, require that the things we put in the table are, say, things that have probability exactly 1. In real life, virtually nothing has probability 1. In an idealisation, many things do. But to argue this way seems to involve using 'idealisation' in an unnatural sense. There is a sense in which, whenever we treat something with non-maximal probability as simply given in a decision problem that we're ignoring, or abstracting away from, some complication. But we aren't idealising. On the contrary, we're modelling the agent as if they were irrationally certain in some things which are merely very very probable.

So it's better to say that any application of decision theory to a real-life problem will involve ignoring certain (counterfactual) logical or metaphysical possibilities in which the decision table is not actually true. But not any old abstraction will do. We can't ignore just anything, at least not if we want a good model. Which abstractions are acceptable? The response I've offered to Dom's challenge suggests an answer to this: we can abstract away from any possibility in which something the agent actually knows is false. I don't have a knock-down argument that this is the best of all possible abstractions, but nor do I know of any alternative answer to the question which abstractions are acceptable which is nearly as plausible.

We might be tempted to say that we can abstract away from anything such that the difference between its probability and 1 doesn't make a difference to the ultimate answer to the decision problem. More carefully, the idea would be that we can have the decision table represent that p iff

p is true and treating $\Pr(p)$ as 1 rather than its actual value doesn't change what the agent should do. I think this is the most plausible story one could tell about decision tables if one didn't like the knowledge first story that I tell. But I also don't think it works, in part because of cases like the following.[2]

Luc is lucky; he's in a casino where they are offering better than fair odds on roulette. Although the chance of winning any bet is $\$1/38$, if Luc bets \$10, and his bet wins, he will win \$400. (That's the only bet on offer.) Luc, like Mireille, is considering betting on 28. As it turns out, 28 won't come up, although since this is a fair roulette wheel, Luc doesn't know this. Luc, like most agents, has a declining marginal utility for money. He currently has \$1,000, and for any amount of money x, Luc gets utility $u(x) = x^{1/2}$ out of having x. So Luc's current utility (from money) is, roughly, 31.622. If he bets and loses, his utility will be, roughly, 31.464. And if he bets and wins, his utility will be, roughly, 37.417. So he stands to gain about 5.794, and to lose about 0.159. So he stands to gain about 36.5 as much as he stands to lose. Since the odds of winning are less than $1/36.5$, his expected utility goes down if he takes the bet, so he shouldn't take it. Of course, if the probability of losing was 1, and not merely $37/38$, he shouldn't take the bet too. Does that mean it is acceptable, in presenting Luc's decision problem, to leave off the table any possibility of him winning, since he won't win, and setting the probability of losing to 1 rather than $37/38$ doesn't change the decision he should make? Of course not; that would horribly misstate the situation Luc finds himself in. It would misrepresent how sensitive Luc's choice is to his utility function, and to the size of the stakes. If Luc's utility function was $u(x) = x^{3/4}$, then he should take the bet. If his utility function is unchanged, but the bet was \$1 against \$40, rather than \$10 against \$400, he should take the bet. Leaving off the possibility of winning hides these facts, and badly misrepresents Luc's situation.

I've argued that the states we can 'leave off' a decision table are the states that the agent knows not to obtain. The argument is largely by elimination. If we can only leave off things that have probability 1, then decision theory would be useless; but it isn't. If we say we can leave off things if setting their probability at 1 is an acceptable idealisation, we need a theory of acceptable idealisations. If this is to be a rival to my theory, the idealisation had better not be it's acceptable to treat anything known as having probability 1. But the most natural alternative idealisation badly

[2]The cases I'll discuss in sections 6.2 and 6.3 also raise problems for this proposal.

misrepresents Luc's case. If we say that what can be left off is not what's known not to obtain, but what is, say, justifiably truly believed not to obtain, we need an argument for why people would naturally use such an unnatural standard. This doesn't even purport to be a conclusive argument, but these considerations point me towards thinking that knowledge determines what we can leave off.

I also cheated a little in making this argument. When I described Mireille in the casino, I made a few explicit comments about her information states. And every time, I said that she knew various propositions. It seemed plausible at the time that this is enough to think those propositions should be incorporated into the table we use to represent her decision. That's some evidence against the idea that more than knowledge, perhaps iterated knowledge or certainty, is needed before we add propositions to the decision table.

If knowledge structures decision tables, then there is a simple argument that Anisa loses knowledge when playing the red-blue game. The following would be a bad table for Anisa to use when deciding what to do.

	$2 + 2 = 4$	$2 + 2 \neq 4$
Red-True	\$50	0
Red-False	0	\$50
Blue-True	\$50	\$50
Blue-False	0	0

4.2 Theoretical Interests

S knows that p only if for any question S is interested in, and any propositons S is interested in, S can rationally answer the queston given those propositions by answering the same (conditional) question conditional on p.

4.3 Knowledge and Closure

Here are two very plausible principles about knowledge, both due to John Hawthorne.

Single Premise Closure If one knows *p* and competently deduces *q* from *p*, thereby coming to believe *q*, while retaining one's knowledge that P, one comes to know that *q*. (Hawthorne, 2005, 43)

Multiple Premise Closure If one knows some premises and competently deduces q from those premises, thereby coming to believe q, while retaining one's knowledge of those premises throughout, one comes to know that q. (Hawthorne, 2005, 43)

Hawthorne endorses the first of these, but has reservations about the second for reasons related to the preface paradox. I'm similarly going to endorse the first and have reservations about the second. But my reasons don't have anything to do with the preface paradox. For reasons I'll discuss in more depth in chapter 10, I think concerns about the preface paradox are over-rated. But I do think we need one more qualification than Hawthorne suggests. I will defend this modification to Hawthorne's second principle.

Relevant Multiple Premise Closure If one knows some premises and competently deduces q from those premises, thereby coming to believe q, while retaining one's knowledge of those premises throughout, and for each conjunction formed from the premises, one has an interest in the question of whether that conjunction is true, one comes to know that q.

4.3.1 Single Premise Closure

It is not trivial to prove that my version of IRT satisfies these closure conditions. One reason for this is that I have not stated a sufficient condition for knowledge. all that I have said is that knowledge is incompatible with a certain kind of caution. So in principle I cannot show that if some conditions obtain then someone knows something. What I can show, is that introducing new conditions linking knowledge with relevant questions as I have done does not introduce new violations of the closure conditions.

But it turns out that even showing this is not completely trivial. Imagine yet another version of the red blue game.[3] In this game, both of the sentences are claims about history that are well supported without being certain. And both of them are supported in the very same way. It turns out to be a little distracting to use concrete examples in this case, so just call the claims A and B. And imagine that the player read both of these claims in the same reliable but not infallible history book, and she knows

[3] This game will resemble the examples that ? and Anderson and Hawthorne (2019b) use to raise doubts about whether pragmatic theories like mine reall do endorse single premise closure.

the book is reliable but not infallible, and she aims to maximise her expected returns. Then all four of the following things are true about the game.

1. Unconditionally, the player is indifferent between playing red-true and playing blue-true.
2. Conditional on A, the player prefers red-true to blue-true, because red-true will certainly return $50 while blue-true is not completely certain to win the money.
3. Conditional on B, the player prefers blue-true to red-true, because blue-true will certainly return $50 while red-true is not completely certain to win the money.
4. Conditional on $A \wedge B$, the player is back to being indifferent between playing red-true and playing blue-true.

From 1, 2 and 3, it follows in my version of IRT that the player does not know either A or B. After all, conditionalising on either one of them changes her answer to a relevant question. The question being, *Which option maximises my expected returns?*, where this is understood as a mention-all question.

But look what happens at point 4. Conditionalising on $A \wedge B$ does not change the answer to that question. So, assuming there is no other reason that the player does not know $A \wedge B$, arguably she does know $A \wedge B$. And that would be absurd; how could she know a conjunction without knowing either conjunct?

Here is how I used to answer this question. Define a technical notion of interest. Say that a person is interested in a conditional question *If p, Q?* if they are interested, in the ordinary sense, in both the true-false question *p?* and they are interested in the question *Q?*. And if conditionalising on a proposition changes (or should change) their answer to any question they are interested in in this technical sense, then they don't know that proposition. This solves the problem because conditionalising on $A \wedge B$ does change their answer to the question *If A, which option maximises expected returns?* on its mention-some reading. So even though 4 is correct, this does pose a problem for closure.

But this is not an entirely satisfactory solution for two reasons. One is that it seems extremely artificial to say that someone is interested in these conditional questions that they have never even formulated. Another is that it is hard to motivate why we should care that conditionalisation changes (or should change) one's answers to these artificial questions.

There was something right about the answer I used to give. It is that we should not just look at whether conditionalisation changes the answers a person gives to questions they are interested in. We should also look at whether it changes things 'under the hood'; whether it changes how they get to that answer. The idea of my old theory was that looking at these artificial questions was a way to indirectly look under the hood. But it is not clear why we should look for these indirect approaches, rather than just looking at what is going on in the player's mind.[4]

So let's look again at the two questions that are relevant. And this time, don't think about what answer the player gives, but about how they get to that answer.

5. Which option maximises expected returns?
6. If $A \wedge B$, which option maximises expected returns?

On the most natural way to understand what the player does, there will be a step in her answer to 5 that has no parallel in her answer to 6.

She will note, and rely on, the fact that she has equally good evidence for A as for B. That is why each option is equally good by her lights. The equality of evidence really matters. If she had read that A in three books, but only one of those books added that B, then the two options would not have the same expected returns. She should check that nothing like this is going on; that the evidence really is equally balanced.

But nothing like this happens in answering 6. In that case, $A \wedge B$ is stipulated to be given. So there is no question about how good the evidence for either is. When answering a question about what to do if a condition obtains, we don't ask how good the evidence for the condition is. We just assume that it holds. So in answering 6, there is no step that acknowledges the equality of the evidence for both A and B.

So in fact the player does not answer the two questions the same way. She ends up with the same conclusion, but she gets there by a different means. And that is enough, I say, to make it a different answer. If she knew $A \wedge B$, she could follow exactly the same steps in answering 5 and 6, but she cannot.

What should we say if she does follow the same steps? If this is irrational, nothing changes, since what matters for knowledge is which questions should be answered the same way, not which questions are answered

[4]Well, unless one is a behaviorist, and so thinks the answers to related questions are all there is to what is going on under the hood. But we should not import that much behaviorism into our epistemology.

the same way. (It does matter for belief, but that is not the current topic.) So I will assume that it is possible for the player to rationally answer both questions the same way. (I will have much more to say about why this is a coherent assumption in chapter 7.)

The way she should answer 6 is to take $A \wedge B$ as given. And hence she will take either option, red-true or blue-true, as being equivalent to just taking $50. And she knows that is the best she can do in the game. So in answering question 6, she will take it as given that both of these options are maximally good.

By hypothesis, she is answering question 5 and question 6 the same way. So she will take it to be part of the setup of question 5 that both options return a sure $50 After all, that is part of the setup of question 6. But if she takes that as given, then conditionalising on either A or B does not change her expected returns. So now claims 2 and 3 are wrong; conditionalising on either conjunct won't make a difference because she treats each conjunct as given.

And that is the totally general case. Assume that someone has competently deduced Y from X, and they know X. So they are entitled to answer the questions *Q?* and *If X, Q?* by the same method. Since the method for the latter takes X as given, so can the method for the former. So they can answer *Q?* taking X as given. What one can appropriately take as given is closed under competent deduction? (Why? Because in the answer to *Q?* that starts with X, you can just go on to derive Y, and then see that it is also a way to answer *If Y, Q?*.) So they can answer *Q?* taking Y as given. So they can answer *Q?* in the same way they answer *If Y, Q?*.

So assuming there is no other reason to deny **Single Premise Closure**, adding a clause about how one may answer questions does not give us a new reason to deny it.

4.4 Puzzles

- Anisa and Blaise - actually in section 1
- Coraline
- And-introduction
- Same source
- Modals (I don't remember what I had in mind here, though maybe it is the stuff about marbles from GBC)
- Miners and doctors (I don't remember what I had in mind here, though maybe it is just 2.2 from KBI)

Chapter 5

Evidence

5.1 A Puzzle About Evidence

Thinknback to the red-blue game at the start of the book. But now consider a version of the game where:

- The red sentence is that two plus two equals four.
- The blue sentence is something that, if known, would be part of the agent's evidence.

I'm going to argue that there are cases where the only rational play is Red-True, but the blue sentence is something we want to say that, ordinarily, the subject knows. And I'll argue that this is a problem for the theory I have described so far. It is not a problem that shows that anything I've said so far is untrue. But it does suggest that what I've said so far is incomplete, and in a key respect unexplanatory.

I have tried so far to argue that belief, rational belief, and knowledge are all interest-relative. And I have tried to tell a story about when they are interest-relative. In the case of knowledge, the story is reasonably simple. One loses knowledge that p when the situation changes in such a way that one is no longer entitled to take p as given in deliberation. But what does it mean to say that one is entitled to take something as given? I haven't given anything like a full theory of this, but the suggestion has been to interpret this on broadly evidentialist lines. In general, one is not entitled to take p as given if the optimal choice, given one's evidence, is different unconditionally to what it is conditional on p. (Exception, as noted in section XXX, one might be rationally following a policy of ignoring certain kinds of evidence. But even in that case, what makes the policy rational

is that one has evidence that costs of ignoring that kind of evidence will, in the long run, be less than the resources one would spend in properly accounting for the evidence. The story is still evidentialist, just not in quite so direct a way as you might have thought.)

But that story doesn't explain when practical considerations might affect what evidence one has. Indeed, it can't explain anything about evidence, since it takes evidence as a given. So if the arguments for the interest-relativity of knowledge can be repurposed to show that evidence too is interest-relative, we have a problem. Since I think they can be repurposed in just this way, there is a problem. The aim of this chapter is to set out just what the problem is, and to suggest a solution to it. I'll start by arguing that evidence is interest-relative, then come back to what the problem is, and how I'll aim to solve it.

Note to self I've included down to 'part of her evidence' in 2.3.4. Refer back to that discussion.

So let's imagine a new player for the red-blue game. Call her Parveen. She is playing the game in a restaurant. It is near her apartment in Ann Arbor, Michigan. Just before the game starts, she notices an old friend, Rahul, across the room. Rahul is someone she knows well, and can ordinarily recognise, but she had no idea he was in town. She thought Rahul was living in Italy. Still, we would ordinarily say that she now knows Rahul is in town; indeed that he is in the restaurant. As evidence for this, note that it would be perfectly acceptable for her to say to someone else, "I saw Rahul here". Now the game starts.

- The red sentence is *Two plus two equals four*.
- The blue sentence is *Rahul is in this restaurant*.

And here is the problem. On the one hand, there is only one rational play for Parveen: Red-True. She hasn't seen Rahul in ages, and she thought he was in Italy. A glimpse of him across a crowded restaurant isn't enough for her to think that *Rahul is in this restaurant* is as likely as *Two plus two equals four*. She might be wrong about Rahul, so she should take the sure money and play Red-True. So playing the red-blue game with these sentences makes it the case that Parveen doesn't know where Rahul is. This is another case where knowledge is interest-relative, and at first glance it doesn't look very different to the other cases we've seen.

But take a second look at the story for why Parveen doesn't know where Rahul is. It can't be just that her evidence makes it certain that two plus

two equals four, but not certain that Rahul is in the restaurant. At least, it can't be that unless it is not part of her evidence that Rahul is in the restaurant. And if evidence is not interest-relative, then I think we should say that it is part of Parveen's evidence that Rahul is in the restaurant. This isn't something she infers; it is a fact about the world she simply appreciated. Ordinarily, it is a starting point for her later deliberations, such as deliberations about whether to walk over to another part of the restaurant to say hi to Rahul. That is, ordinarily it is part of her evidence.

I haven't told you a story about how evidence can be interest-relative. I haven't even started such a story. All the stories I've told you so far about interest-relativity have presupposed that the relevant evidence can be identified, and then we ask what the evidence warrants as circumstances change. That model is by its nature incapable of saying anything about when interests, or practical situations, affect evidence. The model isn't wrong - but it is in a crucial respect incomplete. On the one hand, all models are incomplete. On the other hand, it would be odd to have the model's explanatory ambitions stop somewhere between Anisa's case and Parveen's. That's the kind of explanatory failure that makes one wonder whether you've got the original cases right.

There are two ways out of this problem that I don't want to take, but are notable enough that I want to set them aside explicitly.

One is to say that propositions like *Rahul is in this restaurant* are never part of Parveen's evidence. Perhaps her evidence just consists of things like *I am being appeared to Rahul-like*. Such an approach is problematic for two reasons. The first is that it is subject to all the usual objections to psychological theories of evidence (Williamson, 2007). The second is that we can re-run the argument with the blue sentence being some claim about Parveen's psychological state, and still get the result that the only rational play is Red-True. A retreat to a psychological conception of evidence will only help with this problem if agents are infallible judges of their own psychological states, and that is not in general true (Schwitzgebel, 2008).

Another option is to deny that any explanation is needed here. Perhaps pragmatic effects, like the particular sentences that are chosen for this instance of the red-blue game, mean that Parveen's evidence no longer includes facts about Rahul, and this is a basic epistemic fact without explanation. Now we shouldn't assume that everything relevant to epistemology will have an epistemic explanation. Facts about the way that proteins work in the brain do not have explanations within epistemology, although they are vitally important for there even being a subject matter of epistemology.

So in principle there could be facts around here that ground epistemic explanations without having explanations within epistemology. But in practice things look less rosy. Without an explanation of why Parveen loses evidence, we don't have a theory that makes predictions about how interests affect knowledge. And we don't have a satisfying explanation of why playing Blue-True is irrational for Parveen. And we are forced, as already noted, to draw an implausible distinction between Anisa and Parveen.

We shouldn't be content with simply saying Parveen loses evidence when playing the red-blue game. We should say why this is so. The aim of the rest of this chapter is to tell a story that meets this explanatory desideratum.

5.2 A Simple, but Incomplete, Solution

Let's take a step back and look at the puzzle more abstractly. We have a person S, who has some option O, and it really matters whether or not the value of O, i.e., $V(O)$ is at least x. (I am assuming that Parveen is in the business of maximising utility here. Unlike David in the supermarket from section XXX, she can't simply ignore odd possibilities like that she is wrong about who is in the restaurant.) It is uncontroversial that her evidence includes some background K, and controversial whether it includes some contested proposition p. It is also uncontroversial that $V(O|p) \geq x$, and we're assuming that for any proposition q that is in her evidence, $V(O|q) = V(O)$. That is, we're assuming the relevant values are conditional on evidence. We can capture that last assumption with one big assumption that probably isn't true, but is a harmless idealisation for the purposes of this chapter. Say there is a prior value function V^-, with a similar metaphysical status to the mythical, mystical prior probability function. Then for any choice C, $V(C) = V^-(C|E)$, where E is the evidence the agent has.

Now I can offer a simple, but incomplete, solution. Let p be the proposition that she might or might not know, and the question of whether $V(O) \geq x$ be the only salient one that p is relevant to. Then she knows p only if the following is true:

$$\frac{V^-(O|K)+V^-(O|K\wedge p)}{2} \geq x$$

That is, we work out the value of O with and without the evidence p, and if the average is greater than x, good enough!

That solves the problem of Parveen and Rahul. Parveen's evidence may or may not include that Rahul is in the restaurant. If it does, then Blue-True has a value of $50. If it does not, then Blue-True's value is somewhat lower. Even if the evidence includes that someone who looks a lot like Rahul is in the restaurant, the value of Blue-True might only be $45. Averaging them out, the value is less than $50. But she'd only play Blue-True if it was worthwhile it play it instead of Red-True, which is worth $50. So she shouldn't play Blue-True.

Great! Well, great except for two monumental problems.

The first problem is that what I've said here really only helps with very simple cases, where there is a single decision problem that a single contested proposition is relevant to. There has to be some way to generalise the case to less constrained situations.

The second (and bigger) problem is that the solution is completely ad hoc. Why should the arithmetic mean of these two things have any philosophical significance? Why not the mean of two other things? Why not some other thing, like the geometric mean of them? This looks like a formula plucked out of the air, and there are literally infinitely many other formulae that would do just as well by the one criteria I've laid down so far: imply that Parveen must play red-true.

Pragmatic encroachment starts with a very elegant, very intuitive, principle: you only know the things you can reasonable take to be settled for the purposes of current deliberation. It does not look like any such elegant, intuitive, principles will lead to some theorem about averaging out the value of an option with and without new evidence.

Happily, the two problems have a common solution. But the solution requires a detour into some technical work concerning coordination games.

5.3 The Radical Interpreter

Many philosophical problems can be usefully thought of as games, and hence studied using game theoretic techniques. This is especially when the problems involve interactions of rational agents. Here, for example, is the game table for Newcomb's problem, with the human who is usually the focus of the problem as Row, and the demon as Column.[1]

[1]In these games, Row chooses a row, and Column chooses a column, and that determines the cell that is the outcome of the game. The cells include two numbers. The first is Row's payout, and the second is Column's. The games are non-competitive; the players are

	Predict 1 Box	Predict 2 Boxes
Choose 1 Box	1000, 1	0,0
Choose 2 Boxes	1001, 0	1, 1

This game has a unique Nash equilbrium; the bottom right corner.[2] And that's one way of motivating the view that (a) the game is possible, and (b) the rational move for the human is to choose two boxes.

Let's look at a more complicated game. I'll call it The Interpretation Game. The game has two players. Just like in Newcomb's problem, one of them is a human, the other is a philosophical invention. But in this case the invention is not a demon, but The Radical Interpreter. To know the payouts for the players, we need to know their value function. More colloquially, we need to know their goals.

- The Radical Interpreter assigns mental states to Human in such a way as to predict Human's actions given Human rationality. We'll assume here that evidence is a mental state, so saying what evidence Human has is among Radical Interpreter's tasks. (Indeed, in the game play to come, it will be their primary task.)
- Human acts so as to maximise the expected utility of their action, conditional on the evidence that they have. Human doesn't always know what evidence they have; it depends on what The Radical Interpreter says.

The result is that the game is a coordination game. The Radical Interpreter wants to assign evidence in a way that predicts rational Human action, and Human wants to do what's rational given that assignment of evidence. Coordination games typically have multiple equilibria, and this one is no exception.

Let's make all that (marginally) more concrete. Human is offered a bet on p. If the bet wins, it wins 1 util; if the bet loses, it loses 100 utils. Human's only choice is to Take or Decline the bet. The proposition p, the subject of the bet, is like the claim that Rahul is in the restaurant. It is something that is arguably part of Human's evidence. Unfortunately, it is

simply trying to maximise their own returns, not maximise the difference between their return and the other player's return.

The idea of treating Newcomb's Problem, and similar decision-theoretic problems, as games traces back to ?.

[2] A Nash equilibrium is an outcome of the game where every player does as well as they can given the moves of the other players. Equivalently, it is an outcome where no player can improve their payout by unilaterally defecting from the equilibrium.

also arguable that it is not part of Human's evidence. We will let K be the rest of Human's evidence (apart from p, and things entailed by $K \cup p$), and stipulate that $\Pr(p|K) = 0.9$. Each party now faces a choice.

- The Radical Interpreter has to choose whether p is part of Human's evidence or not.
- Human has to decide whether to Take or Decline the bet.

The Radical Interpreter achieves their goal if human takes the bet iff p is part of their evidence. If p is part of the evidence, then The Radical Interpreter thinks that the bet has positive expected utility, so Human will take it. And if p is not part of the evidence, then The Radical Interpreter thinks that the bet has negative expected utility, so Human will decline it. Either way, The Radical Interpreter wants Human's action to coordinate with theirs. And Human, of course, wants to maximise expected utility. So we get the following table for the game.

	$p \in E$	$p \notin E$
Take the bet	1, 1	-9.1, 0
Decline the bet	0, 0	0, 1

We have, in effect, already covered The Radical Interpreter's payouts. They win in the top-left and lower-right quadrants, and lose otherwise. Human's payouts are only a little trickier. In the bottom row, they are guaranteed 0, since the bet is declined. In the top-left, the bet is a sure winner; their evidence entails it wins. So they get a payout of 1. In the top-right, the bet wins with probability 0.9, so the expected return of taking it is $1 \times 0.9 - 100 \times 0.1 = -9.1$.

There are two Nash equilibria for the game - I've bolded them below.

	$p \in E$	$p \notin E$
Take the bet	**1, 1**	-9.1, 0
Decline the bet	0, 0	**0, 1**

That there are two equilibria to this game should not come as a surprise. It's a formal parallel to the fact that the pragmatic encroachment theory I've developed so far doesn't make a firm prediction about this game. It is consistent with the theory developed so far that Human's evidence includes p, and they should take the bet, or that due to interest-sensitive features of the case, it does not include p, and they should not take the

bet. The aim of this chapter is to supplement that theory with one that, at least most of the time, makes a firm pronouncement about what the evidence is.

But to do that, I need to delve into somewhat more contested areas of game theory. In particular, I need to introduce some work on equilibrium choice. And to do this, it helps to think about a game that is inspired by an example of Rousseau's.

5.4 Motivating Risk-Dominant Equilibria

At an almost maximal level of abstraction, a two player, two option each game looks like this.

	a	*b*
A	r_{11}, c_{11}	r_{12}, c_{12}
B	r_{21}, c_{21}	r_{22}, c_{22}

We're going to focus on games that have the following eight properties:

- $r_{11} > r_{21}$
- $r_{22} > r_{12}$
- $c_{11} > c_{12}$
- $c_{22} > c_{21}$
- $r_{11} > r_{22}$
- $c_{11} \geq c_{22}$
- $\frac{r_{21}+r_{22}}{2} > \frac{r_{11}+r_{12}}{2}$
- $\frac{c_{12}+c_{22}}{2} \geq \frac{c_{11}+c_{21}}{2}$

The first four clauses say that the game has two (strict) Nash equilibria: *Aa* and *Bb*. The fifth and sixth clauses say that the *Aa* equilibria is **Pareto-optimal**: no one prefers the other equilibria to it. In fact it says something a bit stronger: one of the players strictly prefers the *Aa* equilibria, and the other player does not prefer *Bb*. The seventh and eighth clauses say that the *Bb* equilibria is **risk-optimal**.

I'm going to offer an argument from Hans Carlsson and Eric van Damme (1993) for the idea that in these games, rational players will end up at *Bb*. The game that Human and The Radical Interpreter are playing fits these eight conditions, and The Radical Interpreter is perfectly rational, so this will imply that in that game, The Radical Interpreter will say that $p \notin E$, which is what we aimed to show.

Games satisfying these eight inequalities are sometimes called *Stag Hunt* games. There is some flexibility, and some vagueness, in which of the eight inequalities need to be strict, but that level of detail isn't important here. The name comes from a thought experiment in Rousseau's *Discourse on Inequality*.

> They were perfect strangers to foresight, and were so far from troubling themselves about the distant future, that they hardly thought of the morrow. If a deer was to be taken, every one saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs. (Rousseau, 1913, 209–10)

It is rather interesting to think through which real-life situations are best modeled as Stag Hunts, especially in situations where people have thought that the right model was a version of Prisoners' Dilemma. This kind of thought is one way in to appreciating the virtues of Rousseau's political outlook, and especially the idea that social coordination might not require anything like the heavy regulatory presence that, say, Hobbes thought was needed. But that's a story for another day. What I'm going to focus on is why Rousseau was right to think that a 'stranger to foresight', who is just focussing on this game, should take the rabbit.

To make matters a little easier, we'll focus on a very particular instance of Stag Hunt, as shown here. (From here I'm following Carlsson and van Damme very closely; this is their example, with just the labelling slightly altered.)

	a	*b*
A	4, 4	0, 3
B	3, 0	3, 3

At first glance it might seem like *Aa* is the right choice; it produces the best outcome. This isn't like Prisoners Dilemma, where the best collective outcome is dominated. In fact *Aa* is the best outcome for each individual. But it is risky, and Carlsson and van Damme show how to turn that risk into an argument for choosing *Bb*.

Embed this game in what they call a *global game*. We'll start the game

with each player knowing just that they will play a game with the following payout table, with x to be selected at random from a flat distribution over $[-1, 5]$.

	a	b
A	4, 4	0, x
B	x, 0	x, x

Before they play the game, each player will get a noisy signal about the value of x. There will be signals s_R and s_C chosen (independently) from a flat distribution over $[x - 0.25, x + 0.25]$, and shown to Row and Column respectively. So each player will know the value of x to within $\frac{1}{4}$, and know that the other player knows it to within $\frac{1}{4}$ as well. But this is a margin of error model, and in those models there is very little that is common knowledge. That, they argue, makes a huge difference.

In particular, they prove that iterated deletion of strictly dominated strategies (almost) removes all but one strategy pair. (I'll go over the proof of this in the next subsection.) Each player will play A/a if the signal is greater than 2, and B/b otherwise.[3] Surprisingly, this shows that players should play the risk-optimal strategy even when they know the other strategy is Pareto-optimal. When a player gets a signal in $(2, 3.75)$, then they know that $x < 4$, so Bb is the Pareto-optimal equilibrium. But the logic of the global game suggests the risk-dominant equilibrium is what to play.

Carlsson and van Damme go on to show that many of the details of this case don't matter. As long as (a) there is a margin of error in each side's estimation of the payoffs, and (b) every choice is a dominant option in some version of the global game, then iterated deletion of strongly dominant strategies will lead to each player making the risk-dominant choice.

I conclude from that that risk-dominant choices are rational in these games. There is a limit assumption involved here; what's true for games with arbitrarily small margins of error is true for games with no margin of error. (We'll come back to that assumption below.) And since The Radical Interpreter is rational, they will play the strategy that is not eliminated by deleting dominant strategies. That is, they will play the risk-dominant strategy.

In game with Human, the rational (i.e., risk-dominant) strategy for The Radical Interpreter is to say that $p \notin E$. And in the case of Parveen

[3] Strictly speaking, we can't rule out various mixed strategies when the signal is precisely 2, but this makes little difference, since that occurs with probability 0.

and Rahul, rational (i.e., risk-dominant) strategy for The Radical Interpreter is to say that it is not part of Parveen's evidence that Rahul is in the restaurant. And this is an interest-relative theory of evidence; had Parveen been playing a different game, The Radical Interpreter would have said that it is part of Parveen's evidence that Rahul was in the restaurant.

And from this point all the intuitions about the case fall into place. If it is part of Parveen's evidence that Rahul is in the restaurant, then she knows this. Conversely, if she knows it, then The Radical Interpreter would have said it is part of her evidence, so it is part of her evidence. Parveen will perform the action that maximises expected utility given her evidence. And she will lose knowledge when that disposition makes her do things that would be known to be sub-optimal if she didn't lose knowledge.

In short, this model keeps what was good about the pragmatic encroachment theory developed in the previous chapters, while also allowing that evidence can be interest-relative. It does require a slightly more complex theory of rationality than was previously used. Rather than just model rational agents as utility maximisers, they are modelled as playing playing risk-dominant strategies in coordination games. But it turns out that this is little more than assuming that they maximise evidential expected utility, and they expect others (at least perfectly rational abstract others) to do the same, and they expect those others to expect they will maximise expected utility, and so on.

The rest of this section goes into more technical detail about Carlsson and van Damme's example. Readers not interested in these details can skip ahead to the next and last section. In the first subsection I summarise their argument that we only need iterated deletion of strictly dominated strategies to get the result that rational players will play the risk-dominant strategies. In the second subsection I offer a small generalisation of their argument, showing that it still goes through when one of the players gets a precise signal, and the other gets a noisy signal. And I discuss why that is relevant. (In short, the Radical Interpreter doesn't have to deal with noise, and we want the argument to respect that fact.)

5.4.1 The Dominance Argument for Risk-Dominant Equilibria

Two players, Row (or R) and Column (or C) will play a version of the following game.

	a	*b*
A	4, 4	0, *x*

$B \quad x, 0 \quad x, x$

They won't be told what x is, but they will get a noisy signal of x, drawn from an even distribution over $[x - 0.25, x + 0.25]$. Call these signals s_R and s_C. Each player must then choose A, getting either 4 or 0 depending on the other player's choice, or choose B, getting x for sure.

Before getting the signal, the players must choose a strategy. A strategy is a function from signals to choices. Since the higher the signal is, the better it is to play B, we can equate strategies with 'tipping points', where the player plays B if the signal is above the tipping point, and A below the tipping point. Strictly speaking, a tipping point will pick out not a strategy but an equivalence class of strategies, which differ in how they act if the signal is the tipping point. But since that happens with probability 0, the strategies in the equivalence class have the same expected return, and so we won't aim to distinguish them.

Also, strictly speaking, there are strategies that are not tipping points, because they map signals onto probabilities of playing A, where the probability decreases as A rises. I won't discuss these directly, but it isn't too hard to see how these are shown to be suboptimal using the argument that is about to come. It eases exposition to focus on the pure strategies, and to equate these with tipping points. And since my primary aim here is to explain why the result holds, not to simply repeat an already existing proof, I'll mostly ignore these mixed strategies.

Call the tipping points for Row and Column respectively T_R and T_C. Since the game is symmetric, we'll just have to show that in conditions of common knowledge of rationality, $T_R = 2$. It follows by symmetry that $T_C = 2$ as well. And the only rule that will be used is iterated deletion of strictly dominated strategies. That is, neither player will play a strategy where another strategy does better no matter what the opponent chooses, and they won't play strategies where another strategy does better provided the other player does not play a dominated strategy, and they won't play strategies where another strategy does better provided the other player does not play a strategy ruled out by these first two conditions, and so on.

The return to a strategy is uncertain, even given the other player's strategy. But given the strategies of each player, each players' expected return can be computed. And that will be treated as the return to the strategy pair.

Note first that $T_R = 4.25$ strictly dominates any strategy where $T_R = y > 4.25$. If $s_R \in (4.25, y)$, then T_R is guaranteed to return above 4, and the

alternative strategy is guaranteed to return 4. In all other cases, the strategies have the same return. And there is some chance that $s_R \in (4.25, y)$. So we can delete all strategies $T_R = y > 4.25$, and similarly all strategies $T_C = y > 4.25$. By similar reasoning, we can rule out $T_R < -0.25$ and $T_C < -0.25$.

If $s_R \in [-0.75, 4.75]$, then it is equally likely that x is above s_R as it is below it. Indeed, the posterior distribution of x is flat over $[s_R - 0.25, s_R + 0.25]$. From this it follows that the expected return of playing B after seeing signal s_R is just s_R.

Now comes the important step. Assume that we know that $T_C \leq y > 2$. Now consider the expected return of playing A given various values for $s_R > 2$. Given that the lower T_C is, the higher the expected return is of playing A, we'll just work on the simple case where $T_C = y$, realizing that this is an upper bound on the expected return of A given $T_C \leq y$. The expected return of A is 4 times the probability that Column will play a, i.e., 4 times the probability that $s_C < T_C$. Given all the symmetries that have been built into the puzzle, we know that the probability that $s_C < s_R$ is 0.5. So the expected return of playing A is at most 2 if $s_R \geq y$. But the expected return of playing B is, as we showed in the last paragraph, s_R, which is greater than 2. So it is better to play B than A if $s_R \geq y$. And the difference is substantial, so even if s_R is epsilon less than that y, it will still be better to play B. (This is rather hand-wavy, but I'll go over the more rigorous version presently.)

So if $T_C \leq y > 2$ we can prove that T_R should be lower still, because given that assumption it is better to play B even if the signal is just less than y. Repeating this reasoning over and over again pushes us to it being better to play B than A as long as $s_R > 2$. And the same kind of reasoning from the opposite end pushes us to it being better to play A than B as long as $s_R < 2$. So we get $s_R = 2$ as the uniquely rational solution to the game.

Let's make that a touch more rigorous. Assume that $T_C = y$, and s_r is slightly less than y. In particular, we'll assume that $z = y - s_R$ is in $(0, 0.5)$. Then the probability that $s_C < y$ is $0.5 + 2z - 2z^2$. So the expected return of playing A is $2 + 8z - 8z^2$. And the expected return of playing B is, again, s_R. These will be equal when the following is true. (The working out is a tedious but trivial application of the quadratic formula, plus some rearranging.)

$$s_R = y + \frac{\sqrt{145 - 32y} - 9}{16}$$

So if we know that $T_C \geq y$, we know that $T_R \geq y + \frac{\sqrt{145-32y}-9}{16}$, which will be less than y if $y > 2$. And then by symmetry, we know that T_C must be at most as large as that as well. And then we can use that fact to derive a further upper bound on T_R and hence on T_C, and so on. And this will continue until we push both down to 2. It does require quite a number of steps of iterated deletion. Here is the upper bound on the threshold after n rounds of deletion of dominated strategies. (These numbers are precise for the first two rounds, then just to three significant figures after that.)

Round	Upper Bound on Threshold
1	4.250
2	3.875
3	3.599
4	3.378
5	3.195
6	3.041
7	2.910
8	2.798
9	2.701
10	2.617

That is, $T_R = 4.25$ dominates any strategy with a tipping point above 4.25. And $T_R = 3.875$ dominates any strategy with a higher tipping point than that, assuming $T_C \leq 4.25$. And $T_R \approx 3.599$ dominates any strategy with a higher tipping point than that, assuming $T_C \leq 3.875$. And so on.

And similar reasoning shows that at each stage not only are all strategies with higher tipping points dominated, but so are strategies that assign positive probability (whether it is 1 or less than 1), to playing A when the signal is above the 'tipping point'. So this kind of reasoning rules out all mixed strategies (except those that respond probabilistically to $s_R = 2$).

So it has been shown that iterated deletion of dominated strategies will rule out all strategies except the risk-optimal equilibrium. The possibility that x is greater than the maximal return for A is needed to get the iterated dominance going. And the signal to have an error bar to it, so that each round of iteration removes more strategies. But that's all that was needed; the particular values used are irrelevant to the proof.

5.4.2 Making One Signal Precise

The aim of this sub-section is to prove something that Carllson and van Damme did not prove, namely that the analysis of the previous subsection goes through with very little change if one party gets a perfect signal, while the other gets a noisy signal. So I'm going to discuss the game that is just like the game of the previous subsection, but where it is common knowledge that the signal Column gets, s_C, equals x.

Since the game is no longer symmetric, I can't just appeal to the symmetry of the game as frequently as in the previous subsection. But this only slows the proof down, it doesn't stop it.

We can actually rule out slightly more at the first step in this game than in the previous game. Since Column could not be wrong about x, Column knows that if $s_C > 4$ then playing b dominates playing a. So one round of deleting dominated strategies rules out $T_C > 4$, as well as ruling out $T_R > 4.25$.

At any stage, if we know $T_C \leq y > 2$, then $T_R = y$ dominates $T_R > y$. That's because if $s_R \geq y$, and $T_C \leq y$, then the probability that Column will play a (given Row's signal) is less than 0.5. After all, the signal is just as likely to be above x as below it (as long as the signal isn't too close to the extremes). So if s_R is at or above T_C, then it is at least 0.5 likely that $s_C = x$ is at or above T_C. So the expected return of playing A is at most 2. But the expected return of playing B equals the signal, which is greater than 2. So if Row knows $T_C \leq y > 2$, Row also knows it is better to play B if $s_R \geq y$. And that just means that $T_R \leq y$.

Assume now that it is common knowledge that $T_R \leq y$, for some $y > 2$. And assume that $x = s_C$ is just a little less than y. In particular, define $z = y - x$, and assume $z \in (0, 0.25)$. We want to work out the upper bound on the expected return to Column of playing a. (The return of playing b is known, it is x.) The will be highest when T_R is lowest, so assume $T_R \leq y$. Then the probability that Row plays A is $(1 + 2z)/2$. So the expected return of playing a is $2 + 4z$, i.e., $2 + 4(y - x)$. That will be greater than x only when

$$x < \frac{2 + 4y}{5}$$

And so if it is common knowledge that $T_R \leq y$, then it is best for Column to play b unless $x < \frac{2+4y}{5}$. That is, if it is common knowledge that $T_R \leq y$, then T_C must be at most $\frac{2+4y}{5}$.

So now we proceed in a zig-zag fashion. At one stage, we show that T_R must be as low as T_C. At the next, we show that if it has been proven that T_R takes a particular value greater than 2, then T_C must be lower still. And this process will eventually rule out all values for T_R and T_C greater than 2.

This case is crucial to the story of this chapter because The Radical Interpreter probably does not have an error bar in their estimation of the game they are playing. But it turns out the argument for risk-dominant equilibria being the unique solution to interpretation games is consistent with that. As long as one player has a margin of error, each player should play the risk-dominant equilibria.

5.5 Objections and Replies

I'll end this chapter by going over some worries that one might have about the arguments and models that I've used.

Objection: The formal results of the previous section only go through if we assume that the agents do not know precisely what the payoffs are in the game. We shouldn't assume that what holds for arbitrarily small margins of error will hold in the limit, i.e., when they do know the payoffs.

Reply: If pushed, I would defend the use limit assumptions like this to resolve hard cases like Stag Hunt. But I don't strictly need that assumption here, What is needed is that Parveen doesn't know precisely the probability of Rahul being in the restaurant given the rest of her evidence. Given that evidence is not luminous, as Williamson (2000) shows, this is a reasonable assumption. So the margin of error assumption that Carlsson and van Damme make is not, in this case, an assumption that merely makes the math easier; it is built into the case.

Objection: Even if Parveen doesn't know the payoffs precisely, The Radical Interpreter does. The Radical Interpreter is an idealisation, so they can be taken to be ideal.

Reply: It turns out that Carlsson and van Damme's result doesn't require that both parties are ignorant of the precise values of the payoffs. As long as one party doesn't know the exact value of the payoff, the argument goes through. That was the point of the proof in subsection 5.4.2.

Objection: The formal argument requires that in the 'global game' there are values for x that make A the dominant choice. These cases serve as a base step for an inductive argument that follows. But in Parveen's case, there is no such setting for x, so the inductive argument can't get going.

Reply: What matters is that there are values of x such that A is the strictly dominant choice, and Human (or Parveen) doesn't know that they know that they know, etc., that those values are not actual. And that's true in our case. For all Human (or Parveen) knows that they know that they know that they know…, the proposition in question is not part of their evidence under a maximally expansive verdict on The Radical Interpreter's part. So the relevant cases are there in the model, even if for some high value of n they are known^n not to obtain.

Objection: This model is much more complex than the simple motivation for pragmatic encroachment.

Reply: Sadly, this is true. I would like to have a simpler model, but I don't know how to create one. I suspect any such simple model will just be incomplete; it won't say what Parveen's evidence is. In this respect, any simple model will look just like applying tools like Nash equilibria to coordination games. So more complexity will be needed, one way or another. I think paying this price in complexity is worth it overall, but I can see how some people might think otherwise.

Objection: Change the case involving Human so that the bet loses 15 utils if p is false, rather than 100. Now the risk-dominant equilibrium is that Human takes the bet, and The Radical Interpreter says that p is part of Human's evidence. But note that if it was clearly true that p was not part of Human's evidence, then this would still be too risky a situation for them to know p. So whether it is possible that p is part of Human's evidence matters.

Reply: This is all true, and it shows that the view I'm putting forward is incompatible with some programs in epistemology. In particular, it is incompatible with E=K, since the what it takes to be evidence on this story is slightly different from what it takes to be knowledge. I will come back to this point in section ??.

Objection: Carlsson and van Damme discuss one kind of global game. But there are other global games that have different equilibria. For instance, changing the method by which the noisy signal is selected would change the equilibrium of the global game. So this kind of argument can't show that the risk-dominant equilibrium is the one true solution.

Reply: This is somewhat true. There are other ways of embedding the game involving Human and The Radical Interpreter in global games that lead to different outcomes. They are usually somewhat artificial; e.g., by having the signal be systematically biased in one way. But what really matters is the game where the error in Human's knowledge of the payoffs is

determined by their actual epistemic limitations. I think that will lead to something like the model we have here. But it is possible that the final result will differ a bit from what I have here, or (more likely) have some indeterminacy about just how interests interact with evidence and knowledge. The precise details are ultimately less important to me than whether we can provide a motivated story of how interests affect knowledge and evidence that does not presuppose we know what the agent's evidence is. And the method I've outlined here shows that we can do that, even if we end up tinkering a bit with the details.

Chapter 6

Rational Belief

This chapter discusses the role of rational belief in the version of IRT that I defend. It starts by noting that the theory allows for a new kind of Dharmottara case, where a rational, true belief is not actually knowledge. And I argue that it is a good thing it allows this, for once we see the kind of case in question, it is plausible that it is a Dharmottara case. Then offer two arguments, one of them due to Timothy Williamson and the other novel, for the conclusion that it is possible to have rational credence 1 in a proposition without fully believing it. This undermines two prominent theories of the relationship between credence and full belief: that full belief is credence one, and that full belief is credence above some interest-invariant threshold. I'm going to focus primarily on the versions of those views that say that rational full belief is a matter of having rational credences that meet some property. The arguments that credence one is sometimes insufficient for belief are already problems for those views, but they each have a number of independent problems. I'll end the chapter by noting how the view of rational belief that comes out of IRT is immune to those problems.

6.1 Atomism about Rational Belief

In chapter 3 I argued for two individually necessary and jointly sufficient conditions for belief.[1] They are

1. In some possible decision problem, p is taken for granted.

[1] This section is based on material from my (2012, sect. 3.1).

2. For every question the agent is interested in, the agent answers the question the same way (i.e., giving the same answer for the same reasons) whether the question is asked unconditionally or conditional on p.

At this point one might think that offering a theory of rational belief would be easy. It is rational to believe p just in case it is rational to satisfy these conditions. Unfortunately, this nice thought can't be right. It can be irrational to satisfy these conditions while rationally believing p.

Coraline is like Anisa and Chamari, in that she has read a reliable book saying that the Battle of Agincourt was in 1415. And she now believes that the Battle of Agincourt was indeed in 1415, for the very good reason that she read it in a reliable book.

In front of her is a sealed envelope, and inside the envelope a number is written on a slip of paper. Let X denote that number, non-rigidly. (So when I say Coraline believes $X = x$, it means she believes that the number written on the slip of paper is x, where x rigidly denotes some number.) Coraline is offered the following bet:

- If she declines the bet, nothing happens.
- If she accepts the bet, and the Battle of Agincourt was in 1415, she wins \$1.
- If she accepts the bet, and the Battle of Agincourt was not in 1415, she loses X dollars.

For some reason, Coraline is convinced that $X = 10$. This is very strange, since she was shown the slip of paper just a few minutes ago, and it clearly showed that $X = 1,000,000,000$. Coraline wouldn't bet on when the Battle of Agincourt was at odds of a billion to one. But she would take that bet at 10 to 1, which is what she thinks she is faced with. Indeed, she doesn't even conceptualise it as a bet; it's a free dollar she thinks. Right now, she is disposed to treat the date of the battle as a given. She is disposed to lose this disposition should a very long odds bet appear to depend on it. But she doesn't believe she is facing such a bet.

So Coraline accepts the bet; she thinks it is a free dollar. And that's when the battle took place, so she wins the dollar. All's well that end's well. But it was a really wildly irrational bet to take. You shouldn't bet at those odds on something you remember from a history book. Neither memory nor history books are that reliable. Coraline was not rational to treat the questions *Should I take this bet?*, and *Conditional on the Battle of Agincourt*

being in 1415, should I take this bet? the same way. Her treating them the same way was fortunate - she won a dollar - but irrational.

Yet it seems odd to say that Coraline's belief about the Battle of Agincourt was irrational. What was irrational was her belief about the envelope, not her belief about the battle. To say that a particular disposition was irrational is to make a holistic assessment of the person with the disposition. But whether a belief is rational or not is, relatively speaking, atomistic.

That suggests the following condition on rational belief.

> S's belief that *p* is irrational if
>
> 1. S irrationally has one of the dispositions that is characteristic of belief that *p*; and
> 2. What explains S having a disposition that is irrational in that way is her attitudes towards *p*, not (solely) her attitudes towards other propositions, or her skills in practical reasoning.

In "Knowledge, Bets and Interests" I gave a similar theory about these cases - I said that S's belief that *p* was irrational if the irrational dispositions were caused by an irrationally high credence in *p*. I mean this account to be ever so slightly more general. I'll come back to that below, because first I want to spell out the second clause.

Intuitively, Coraline's irrational acceptance of the belief is explained by her (irrational) belief about *X*, not her (rational) belief about the Battle of Agincourt. We can take this notion of explanation as a primitive if we like; it's in no worse philosophical shape than other notions we take as a primitive. But it is possible to spell it out a little more.

Coraline has a pattern of irrational dispositions related to the envelope. If you offer her \$50 or *X* dollars, she'll take the \$50. If you change the bet so it isn't about Agincourt, but is instead about any other thing she has excellent but not quite conclusive evidence for, she'll still take the bet.

On the other hand, she does not have a pattern of irrational dispositions related to the Battle of Agincourt. She has this one, but if you change the payouts so they are not related to this particular envelope, then for all we have said so far, she won't do anything irrational.

That difference in patterns matters. We know that it's the beliefs about the envelope, and not the beliefs about the battle, that are explanatory because of this pattern. We could try and create a reductive analysis of explanation in clause 2 using facts about patterns, like the way Lewis tries

to create a reductive analysis of causation using similar facts about patterns in "Causation as Influence" (Lewis, 2004). But doing so would invariably run up against edge cases that would be more trouble to resolve than they are worth.

That's because there are ever so many ways in which someone could have an irrational disposition about any particular case. We can imagine Coraline having a rational belief about the envelope, but still taking the bet because of any of the following reasons:

- It has been her life goal to lose a billion dollars in a day, so taking the bet strictly dominates not taking it.
- She believes (irrationally) that anyone who loses a billion dollars in a day goes to heaven, and she (rationally) values heaven above any monetary amount.
- She consistently makes reasoning errors about billions, so the prospect of losing a billion dollars rarely triggers an awareness that she should reconsider things she normally takes for granted.

The last one of these is especially interesting. The picture of rational agency I have in the background here owes a lot to the notion of epistemic vigilance, as developed by Dan Sperber and co-authors (Sperber et al., 2010). The rational agent will have all these beliefs in their head that they will drop when the costs of being wrong about them are too high, or the costs of re-opening inquiry into them are too low. They can't reason, at least in any conscious way, about whether to drop these beliefs, because to do that is, in some sense, to call the belief into doubt. And what's at issue is whether they should call the belief into doubt. So what they need is some kind of disposition to replace a belief that p with an attitude that p is highly probable, and this disposition should correlate with the cases where taking p for granted will not maximise expected utility. This disposition will be a kind of vigilance. As Sperber et al show, we need some notion of vigilance to explain a lot of different aspects of epistemic evaluation, and I think it can be usefully pressed into service here.[2]

But if you need something like vigilance, then you have to allow that vigilance might fail. And maybe some irrational dispositions can be traced to that failure, and not to any propositional attitude the decider has. For example, if Coraline systematically fails to be vigilant when exactly one

[2]Kenneth Boyd (2016) suggests a somewhat similar role for vigilance in the course of defending an interest-invariant epistemic theory. Obviously I don't agree with his conclusions, but my use of Sperber's work does echo his.

billion dollars is at stake, then we might want to say that her belief in p is still rational, and she is practically, rather than theoretically, irrational. (Why could this happen? Perhaps she thinks of Dr Evil every time she hears the phrase "One billion dollars", and this distractor prevents her normally reliable skill of being vigilant from kicking in.)

If one tries to turn the vague talk of patterns of bets involving one proposition or another into a reductive analysis of when one particular belief is irrational, one will inevitably run into hard cases where a decider has multiple failures. We can't say that what makes Coraline's belief about the envelope, and not her belief about the battle, irrational is that if you replaced the envelope, she would invariably have a rational disposition. After all, she might have some other irrational belief about whatever we replace the envelope with. Or she might have some failure of practical reasoning, like a vigilance failure. Any kind of universal claim, like that it is only bets about the envelope that she gets wrong, won't do the job we need.

In "Knowledge, Bets and Interests", I tried to use the machinery of credences to make something like this point. The idea was that Coraline's belief in p was rational because her belief just was her high credence in p, and that credence was rational. I still think that's approximately right, but it can't be the full story.

For one thing, beliefs and credences aren't as closely connected metaphysically as this suggests. To have a belief in p isn't just to have a high credence, it's to be disposed to let p play a certain role. (This will become important in the next two sections.)

For another thing, it is hard to identify precisely what a credence is in the case of an irrational agent. The usual ways we identify credences, via betting dispositions or representation theorems, assume away all irrationality. But an irrational person might still have some rational beliefs.

Attempts to generalise accounts of credences so that they cover the irrational person will end up saying something like what I've said about patterns. What it is to have credence 0.6 in p isn't to have a set of preferences that satisfies all the presuppositions of such and such a representation theorem, and that theorem to say that one can be represented by a probability function Pr and a utility function U such that $\Pr(p) = 0.6$. That can't be right because some people will, intuitively, have credence about 0.6 in p while not uniformly conforming to these constraints. But what makes them intuitive cases of credence roughly 0.6 in p is that generally they behave like the perfectly rational person with credence 0.6 in p, and

most of the exceptions are explained by other features of their cognitive system other than their attitude to *p*.

In other words, we don't have a full theory of credences for irrational beings right now, and when we get one, it won't be much simpler than the theory in terms of patterns and explanations I've offered here. So it's best for now to just understand belief in terms of a pattern of dispositions, and say that the belief is rational just in case that pattern is rational. And that might mean that on some occasions *p*-related activity is irrational even though the pattern of *p*-related activity is a rational pattern. Any given action, like any thing whatsoever, can be classified in any number of ways. What matters here is what explains the irrationality of a particular irrational act, and that will be a matter of which patterns of irrational dispositions the actor has.

However we explain Coraline's belief, the upshot is that she has a rational, true belief that is not knowledge. This is a novel kind of Dharmottara case. (Or Gettier case for folks who prefer that nomenclature.) It's not the exact kind of case that Dharmottara originally described. Coraline doesn't infer anything about the Battle of Agincourt from a false belief. But it's a mistake to think that the class of rational, true beliefs that are not knowledge form a natural kind. In general, negatively defined classes are disjunctive; there are ever so many ways to not have a property. An upshot of this discussion of Coraline is that there is one more kind of Dharmottara case than was previously recognised. But as, for example, Williamson (2013) and Nagel (2013) have shown, we have independent reason for thinking this is a very disjunctive class. So the fact that it doesn't look anything like Dharmottara's example shouldn't make us doubt it is a rational, true belief that is not knowledge.

6.2 Coin Puzzles

So rational belief is not identical to rationally having the dispositions that constitute belief. But nor is rational belief a matter of rational high credence. I'll offer two arguments for this, one in this section and one in the next.

The point of these two sections is as much metaphysical as it is normative. I'm interested in arguing against the 'Lockean' thesis that to believe *p* just is to have a high credence in *p*. Normally, this threshold of high enough belief for credence is taken to be interest-invariant, so this is a rival to IRT. But there is some variation in the literature about whether

the phrase *The Lockean Thesis* refers to a metaphysical claim, belief is high credence, or a normative claim, rational belief is rational high credence. Since everyone who accepts the metaphysical claim also accepts the normative claim, and usually takes it to be a consequence of the metaphysical claim, arguing against the normative claim is a way of arguing against the metaphysical claim.

The first puzzle for this Lockean view comes from an argument that Timothy Williamson (2007) made about certain kinds of infinitary events. A fair coin is about to be tossed. It will be tossed repeatedly until it lands heads twice. The coin tosses will get faster and faster, so even if there is an infinite sequence of tosses, it will finish in a finite time. (This isn't physically realistic, but this need not detain us. All that will really matter for the example is that someone could believe this will happen, and that's physically possible.)

Consider the following three propositions

(A) At least one of the coin tosses will land either heads or tails.
(B) At least one of the coin tosses will land heads.
(C) At least one of the coin tosses after the first toss will land heads.

So if the first coin toss lands heads, and the rest land tails, (B) is true and (C) is false.

Now consider a few versions of the Red-Blue game (perhaps played by someone who takes this to be a realistic scenario). In the first instance, the red sentence says that (B) is true, and the blue sentence says that (C) is true. In the second instance, the red sentence says that (A) is true, and the blue sentence says that (B) is true. In both cases, it seems that the unique rational play is Red-True. But it's really hard to explain this in a way consistent with the Lockean view.

Williamson argues that we have good reason to believe that the probability of all three sentences is 1. For (B) to be false requires (C) to be false, and for one more coin flip to land tails. So the probability that (B) is false is one-half the probability that (C) is false. But we also have good reason to believe that the probabilities of (B) and (C) are the same. In both cases, they are false if a countable infinity of coin flips lands tails. Assuming that the probability of some sequence having a property supervenes on the probabilities of individual events in that sequence (conditional, perhaps, on other events in the sequence), it follows that the probabilities of (B) and (C) are identical. And the only way for the probability that (B) is false to be half the probability that (C) is false, while (B) and (C) have

the same probability, is for both of them to have probability 1. Since the probability of (A) is at least as high as the probability of (B) (since it is true whenever (B) is true, but not conversely), it follows that the probability of all three is 1.

But since betting on (A) weakly dominates betting on (B), and betting on (B) weakly dominates betting on (C), we shouldn't have the same attitudes towards bets on these three propositions. Given a choice between betting on (B) and betting on (C), we should prefer to bet on (B) since there is no way that could make us worse off, and some way it could make us better off. Given that choice, we should prefer to bet on (B) (i.e., play Red-True when (B) and (C) are expressed by the red and blue sentences), because it might be that (B) is true and (C) false.

Assume (something the Lockean may not wish to acknowledge) that to say something might be the case is to reject believing its negation. Then a rational person faced with these choices will not believe *Either (B) is false or (C) is true*; they will take its negation to be possible. But that proposition is at least as probable as (C), so it too has probability 1. So probability 1 does not suffice for belief.

This is a real problem for the Lockean - no probability suffices for belief, not even probability 1. It's also, of course, a problem for the view that belief is probability 1 that I discussed back in section 6.5. The next section discusses another way in which these two views are problematic.

6.3 Playing Games

Some people might be nervous about resting too much weight on infinitary exaples like the coin sequence. So I'll show how the same puzzle arises in a simple, and finite, game.[3] The game itself is a nice illustration of how a number of distinct solution concepts in game theory come apart. (Indeed, the use I'll make of it isn't a million miles from the use that Kohlberg and Mertens (1986) make of it.) To set the problem up, I need to say a few words about how I think of game theory. This won't be at all original - most of what I say is taken from important works by Robert Stalnaker (1994; 1996; 1998; 1999). But the underlying philosophical points are important, and it is easy to get confused about them. (At least, I used to get these points all wrong, and that's got to be evidence they are easy to get confused about, right?) So I'll set the basic points slowly, and

[3] This is based on material in my (2014, sect. 1).

then circle back to the puzzle for the Lockeans.[4]

Start with a simple decision problem, where the agent has a choice between two acts A_1 and A_2, and there are two possible states of the world, S_1 and S_2, and the agent knows the payouts for each act-state pair are given by the following table.

	S_1	S_2
A_1	4	0
A_2	1	1

What to do? I hope you share the intuition that it is radically underdetermined by the information I've given you so far. If S_2 is much more probable than S_1, then A_2 should be chosen; otherwise A_1 should be chosen. But I haven't said anything about the relative probability of those two states. Now compare that to a simple game. Row has two choices, which I'll call A_1 and A_2. Column also has two choices, which I'll call S_1 and S_2. It is common knowledge that each player is rational, and that the payouts for the pairs of choices are given in the following table. (As always, Row's payouts are given first.)

	S_1	S_2
A_1	4, 0	0, 1
A_2	1, 0	1, 1

What should Row do? This one is easy. Column gets 1 for sure if she plays S_2, and 0 for sure if she plays S_1. So she'll play S_2. And given that she's playing S_2, it is best for Row to play A_2.

You probably noticed that the game is just a version of the decision problem from a couple of paragraphs ago. The relevant states of the world are choices of Column. But that's fine; the layout of that decision problem was neutral on what constituted the states S_1 and S_2. Note that the game can be solved without explicitly saying anything about probabilities. What is added to the (unsolvable) decision-theoretic problem is not information about probabilities, but information about Column's payouts, and the fact that Column is rational. Those facts imply something about Column's play, namely that she would play S_2. And that settles what Row should

[4]I'm grateful to the participants in a game theory seminar at Arché in 2011, especially Josh Dever and Levi Spectre, for very helpful discussions that helped me see through my previous confusions.

do.

There's something quite general about this example. What's distinctive about game theory isn't that it involves any special kinds of decision making. Once we get the probabilities of each move by the other player, what's left is (mostly) expected utility maximisation.[5] The distinctive thing about game theory is that the probabilities aren't specified in the setup of the game; rather, they are solved for. Apart from special cases, such as where one option strictly dominates another, not much can be said about a decision problem with unspecified probabilities. But a lot can be said about games where the setup of the game doesn't specify the probabilities, because it is possible to solve for the probabilities given the information that is provided.

This way of thinking about games makes the description of game theory as 'interactive epistemology' (Aumann, 1999) rather apt. The theorist's work is to solve for what a rational agent should think other rational agents in the game should do. From this perspective, it isn't surprising that game theory will make heavy use of equilibrium concepts. In solving a game, we must deploy a theory of rationality, and attribute that theory to rational actors in the game itself. In effect, we are treating rationality as something of an unknown, but one that occurs in every equation we have to work with. Not surprisingly, there are going to be multiple solutions to the puzzles we face.

This way of thinking lends itself to an epistemological interpretation of one of the most puzzling concepts in game theory, the mixed strategy. The most important solution concept in modern game theory is the Nash equilibrium. A set of moves is a Nash equilibrium if no player can improve their outcome by deviating from the equilibrium, conditional on no other player deviating. In many simple games, the only Nash equilibria involve mixed strategies. Here's one simple example.

	S_1	S_2
A_1	0, 1	10, 0
A_2	9, 0	-1, 1

This game is reminiscent of some puzzles that have been much discussed in the decision theory literature, namely asymmetric Death in Damascus

[5] The qualification is because weak dominance reasoning cannot be construed as orthodox expected utility maximisation. We saw that in the coins case, and it will become important again here. It is possible to model weak dominance reasoning using non-standard probabilities, as in ?, but that introduces new complications.

puzzles [Richter (1984);] . Here Column wants herself and Row to make the 'same' choice, i.e., A_1 and S_1 or A_2 and S_2. She gets 1 if they do, 0 otherwise. And Row wants them to make different choices, and gets 10 if they do. Row also dislikes playing A_2, and this costs her 1 whatever else happens. It isn't too hard to prove that the only Nash equilibrium for this game is that Row plays a mixed strategy playing both A_1 and A_2 with probability 1/2, while Column plays the mixed strategy that gives S_1 probability 11/20, and S_2 with probability 9/20.

Now what is a mixed strategy? It is easy enough to take away form the standard game theory textbooks a **metaphysical** interpretation of what a mixed strategy is. Here, for instance, is the paragraph introducing mixed strategies in Dixit and Skeath's *Games of Strategy*.

> When players choose to act unsystematically, they pick from among their pure strategies in some random way …We call a random mixture between these two pure strategies a mixed strategy. (Dixit and Skeath, 2004, 186)

Dixit and Skeath are saying that it is definitive of a mixed strategy that players use some kind of randomisation device to pick their plays on any particular run of a game. That is, the probabilities in a mixed strategy must be in the world; they must go into the players' choice of play. That's one way, the paradigm way really, that we can think of mixed strategies metaphysically.

But the understanding of game theory as interactive epistemology naturally suggests an **epistemological** interpretation of mixed strategies.

> One could easily … [model players] … turning the choice over to a randomizing device, but while it might be harmless to permit this, players satisfying the cognitive idealizations that game theory and decision theory make could have no motive for playing a mixed strategy. So how are we to understand Nash equilibrium in model theoretic terms as a solution concept? We should follow the suggestion of Bayesian game theorists, interpreting mixed strategy profiles as representations, not of players' choices, but of their beliefs. (Stalnaker, 1994, 57-8)

One nice advantage of the epistemological interpretation, as noted by Binmore (2007, 185) is that we don't require players to have n-sided dice in

their satchels, for every n, every time they play a game.[6] But another advantage is that it lets us make sense of the difference between playing a pure strategy and playing a mixed strategy where one of the 'parts' of the mixture is played with probability one.

With that in mind, consider the below game, which I'll call Up-Down.[7] Informally, in this game A and B must each play a card with an arrow pointing up, or a card with an arrow pointing down. I will capitalise A's moves, i.e., A can play UP or DOWN, and italicise B's moves, i.e., B can play *up* or *down*. If at least one player plays a card with an arrow facing up, each player gets \$1. If two cards with arrows facing down are played, each gets nothing. Each cares just about their own wealth, so getting \$1 is worth 1 util. All of this is common knowledge. More formally, here is the game table, with A on the row and B on the column.

	up	*down*
UP	1, 1	1, 1
DOWN	1, 1	0, 0

When I write game tables like this, I mean that the players know that these are the payouts, that the players know the other players to be rational, and these pieces of knowledge are common knowledge to at least as many iterations as needed to solve the game. (I assume here that in solving the game, it is legitimate to assume that if a player knows that one option will do better than another, they have conclusive reason to reject the latter option. This is completely standard in game theory, though somewhat controversial in philosophy.) With that in mind, let's think about how the agents should approach this game.

I'm going to make one big simplifying assumption at first. I'll relax this later, but it will help the discussion to start with this assumption. This assumption is that the doctrine of Uniqueness applies here; there is precisely one rational credence to have in any salient proposition about how the game will play. Some philosophers think that Uniqueness always holds

[6] It is worse than if some games have the only equilibria involving mixed strategies with irrational probabilities. And it might be noted that Binmore's introduction of mixed strategies, on page 44 of his (2007), sounds much more like the metaphysical interpretation. But I think the later discussion is meant to indicate that this is just a heuristic introduction; the epistemological interpretation is the correct one.

[7] In earlier work I'd called it Red-Green, but this is too easily confused with the Red-Blue game that plays such an important role in chapter 2.

(White, 2005). I join with those such as North (2010) and Schoenfield (2013) who don't. But it does seem like Uniqueness might often hold; there might often be a right answer to a particular problem. Anyway, I'm going to start by assuming that it does hold here.

The first thing to note about the game is that it is symmetric. So the probability of *A* playing UP should be the same as the probability of *B* playing *up*, since *A* and *B* face exactly the same problem. Call this common probability x. If $x < 1$, we get a quick contradiction. The expected value, to Row, of UP, is 1. Indeed, the known value of UP is 1. If the probability of *up* is x, then the expected value of UP is x. So if $x < 1$, and Row is rational, she'll definitely play UP. But that's inconsistent with the claim that $x < 1$, since that means that it isn't definite that Row will play UP.

So we can conclude that $x = 1$. Does that mean we can know that Row will play UP? No. Assume we could conclude that. Whatever reason we would have for concluding that would be a reason for any rational person to conclude that Column will play *up*. Since any rational person can conclude this, Row can conclude it. So Row knows that she'll get 1 whether she plays UP or DOWN. But then she should be indifferent between playing UP and DOWN. And if we know she's indifferent between playing UP and DOWN, and our only evidence for what she'll play is that she's a rational player who'll maximise her returns, then we can't be in a position to know she'll play UP.

For the rest of this ssection I want to reply to one objection, and weaken an assumption I made earlier. The objection is that I'm wrong to assume that agents will only maximise expected utility. They may have tie-breaker rules, and those rules might undermine the arguments I gave above. The assumption is that there's a uniquely rational credence to have in any given situation.

I argued that if we knew that *A* would play UP, we could show that *A* had no reason to play UP. But actually what we showed was that the expected utility of playing UP would be the same as playing DOWN. Perhaps *A* has a reason to play UP, namely that UP weakly dominates DOWN. After all, there's one possibility on the table where UP does better than DOWN, and none where RED does better. And perhaps that's a reason, even if it isn't a reason that expected utility considerations are sensitive to.

Now I don't want to insist on expected utility maximisation as the only rule for rational decision making. Sometimes, I think some kind of tie-breaker procedure is part of rationality. In the papers by Stalnaker I

mentioned above, he often appeals to this kind of weak dominance reasoning to resolve various hard cases. But I don't think weak dominance provides a reason to play UP in this particular case. When Stalnaker says that agents should use weak dominance reasoning, it is always in the context of games where the agents' attitude towards the game matrix is different to their attitude towards each other. One case that Stalnaker discusses in detail is where the game table is common knowledge, but there is merely common (justified, true) belief in common rationality. Given such a difference in attitudes, it does seem there's a good sense in which the most salient departure from equilibrium will be one in which the players end up somewhere else on the table. And given that, weak dominance reasoning seems appropriate.

But that's not what we've got here. Assuming that rationality requires playing UP/*up*, the players know we'll end up in the top left corner of the table. There's no chance that we'll end up elsewhere. Or, perhaps better, there is just as much chance we'll end up 'off the table', as that we'll end up in a non-equilibrium point on the table. To make this more vivid, consider the 'possibility' that *B* will play *across*, and if *B* plays *across*, *A* will receive 2 if she plays DOWN, and -1 if she plays UP. Well hold on, you might think, didn't I say that *up* and *down* were the only options, and this was common knowledge? Well, yes, I did, but if the exercise is to consider what would happen if something the agent knows to be true doesn't obtain, then the possibility that one agent will play blue certainly seems like one worth considering. It is, after all, a metaphysical possibility. And if we take it seriously, then it isn't true that under any possible play of the game, UP does better than DOWN.

We can put this as a dilemma. Assume, for reductio, that UP/*up* is the only rational play. Then if we restrict our attention to possibilities that are epistemically open to *A*, then UP does just as well as DOWN; they both get 1 in every possibility. If we allow possibilities that are epistemically closed to *A*, then the possibility where *B* plays *blue* is just as relevant as the possibility that *B* is irrational. After all, we stipulated that this is a case where rationality is common knowledge. In neither case does the weak dominance reasoning get any purchase.

With that in mind, we can see why we don't need the assumption of Uniqueness. Let's play through how a failure of Uniqueness could undermine the argument. Assume, again for reductio, that we have credence $\varepsilon > 0$ that *A* will play DOWN. Since *A* maximises expected utility, that means *A* must have credence 1 that *B* will play *up*. But this is already

odd. Even if you think people can have different reactions to the same evidence, it is odd to think that one rational agent could regard a possibility as infinitely less likely than another, given isomorphic evidence. And that's not all of the problems. Even if A has credence 1 that B will play *up*, it isn't obvious that playing UP is rational. After all, relative to the space of epistemic possibilities, UP weakly dominates DOWN. Remember that we're no longer assuming that it can be known what A or B will play. So even without Uniqueness, there are two reasons to think that it is wrong to have credence $\varepsilon > 0$ that A will play DOWN. So we've still shown that credence 1 doesn't imply knowledge, and since the proof is known to us, and full belief is incompatible with knowing that you can't know, this is a case where credence 1 doesn't imply full belief. So whether A plays UP, like whether the coin will ever land tails, is a case where belief comes apart from high credence, even if by high credence we literally mean credence one. This is a problem for the Lockean, and, like Williamson's coin, it is also a problem for the view that belief is credence one.

6.4 Puzzles for Lockeans

The phrase "the Lockean theory of belief" is sometimes used for a metaphysical claim, and sometimes for a normative claim. The two claims are closely related. The metaphysical claim is that what it is to believe p is to have a credence in p above some threshold t. The normative claim is that one should believe p if and only if one's (rational) credence in p is above some threshold t. Crucially, this threshold is meant to be interest-invariant. If we restrict our attention to rational agents, then there won't be a huge difference between the metaphysical and normative versions of the Lockean theory. At the very least, one of them will be extensionally adequate if and only if the other is. So that's what I'll do in this section. I'll look at various rational agents, and show what puzzles arise if we assume that belief goes with high credence.

I've already mentioned two classes of puzzles, those to do with infinite sequences of coin tosses and those to do with weak dominance in games. But there are other puzzles that apply especially to the Lockean.

6.4.1 Arbitrariness

The first problem for the Lockeans, and in a way the deepest, is that it makes the boundary between belief and non-belief arbitrary. This is a

point that was well made some years ago now by Robert Stalnaker (1984, 91). Unless these numbers are made salient by the environment, there is no special difference between believing p to degree 0.9876 and believing it to degree 0.9875. But if t is 0.98755, this will be *the difference* between believing p and not believing it, which is an important difference.

The usual response to this, as found in Foley (1993, Ch. 4) and Hunter (1996) and Lee (2017a), is to say that the boundary is vague. Now we might respond to this by noting that this only helps on an implausible theory of vagueness. On epistemicist theories, or supervaluationist theories, or on my preferred comparative truth theory (Weatherson (2005b)), there will still be an arbitrary point which marks the difference between belief and non-belief. This won't be the case on various kinds of degree of truth theories. But, as Williamson (1994) pointed out, those are theories on which contradictions end up being half-true. And if saving the Lockean theory requires that we give up on the idea that contradictions are simply false, it is hard to see how it is worth the price.

But a better response is to think about what it means to say that the belief/non-belief boundary is a vague point on a scale. We know plenty of terms where the boundary is a vague point on a scale. Comparative adjectives are typically like that. Whether a day is hot depends on whether it is above some vague point on a temperature scale, for example. But here's the thing about these vague terms - they don't enter into lawlike generalisations. (At least in a non-trivial way. Hot days are 24 hours long, and that's a law, but not one that hotness has a particular role in grounding.) The laws involve the scale; the most you can say using the vague term is some kind of generic. For instance, you can say that hot days are exhausting, or that electricity use is higher on hot days. But these are generics, and the interesting law-like claims will involve degrees of heat, not the hot/non-hot binary.

It's a fairly central presupposition of this book that belief is not like that. Belief plays a key role in all sorts of non-trivial lawlike generalisations. Folk psychology is full of such lawlike generalisations. We're doing social science here, so the laws in question are hardly exceptionless. But they are counterfactually resilient, and explanatorily deep, and not just generics that are best explained using the underlying scale.

Of course, the Lockean doesn't believe that these generalisations of folk psychology are anything more than generics, so this is a somewhat question-begging argument. But if you're not antecedently disposed to give up on folk psychology, or reduce it to the status of a bunch of help-

ful generics, it's worth seeing how striking the Lockean view here is. So consider a generalisation like the following.

- If someone wants an outcome O, and they believe that doing X is the only way to get O, and they believe that doing X will neither incur any costs that are large in comparison to how good O is, nor prevent them being able to do something that brings about some other outcome that is comparatively good, then they will do X.

This isn't a universal - some people are just practically irrational. But it's stronger than just a generic claim about high temperatures. Or so I say. But the Lockean does not say this; they say that this has widespread counterexamples, and when it is true, it is a relatively superficial truth whose explanatory force is entirely derived from deeper truths about credences.

The Lockean, for instance, thinks that someone in Blaise's situation satisfies all the antecedents and qualifications in the principle. They want the child to have a moment of happiness. They believe (i.e., have a very high credence that) taking the bet will bring about this outcome, will have no costs at all, and will not prevent them doing anything else. Yet they will not think that people in Blaise's situation will generally take the bet, or that it would be rational for them to take the bet, or that taking the bet is explained by these high credences.

That's what's bad about making the belief/non-belief distinction arbitrary. It means that generalisations about belief are going to be not particularly explanatory, and are going to have systematic (and highly rational) exceptions. We should expect more out of a theory of belief.

6.4.2 Correctness

I've talked about this one a bit in subsection 3.6.1, so I'll be brief here. Beliefs have correctness conditions. To believe p when p is false is to make a mistake. That might be an excusable mistake, or even a rational mistake, but it is a mistake. On the other hand, having an arbitrarily high credence in p when p turns out to be false is not a mistake. So having high credence in p is not the same as believing p.

Matthew Lee (2017b) argues that the versions of this argument by Ross and Schroeder (2014) and Fantl and McGrath (2009) are incomplete because they don't provide a conclusive case for the premise that having a high credence in a falsehood is not a mistake. But this gap can be plugged. Imagine a scientist, call her Marie, who knows the correct

theory of chance for a given situation. She knows that the chance of p obtaining is 0.999. (If you think $t > 0.999$, just increase this number, and change the resulting dialogue accordingly.) And her credence in p is 0.999, because her credences track what she knows about chances. She has the following exchange with an assistant.

> ASSISTANT: Will p happen?
> MARIE: Probably. It might not, but there is only a one in a thousand chance of that. So p will probably happen.

To their surprise, p does not happen. But Marie did not make any kind of mistake here. Indeed, her answer to assistant's question was exactly right. But if the Lockean theory of belief is right, and false beliefs are mistakes, then Marie did make a mistake. So the Lockean theory of belief is not right.

6.4.3 Moorean Paradoxes

The Lockean says other strange things about Marie. By hypothesis, she believes that p will obtain. Yet she certainly seems sincere when she says it might not happen. So she believes both p and it might not be that p. This looks like a Moore-paradoxical utterance, yet in context it seems completely banal.

The same thing goes for Chamira. Does she believe the Battle of Agincourt was in 1415? Yes, say the Lockeans. Does she also believe that it might not have been in 1415? Yes, say the Lockeans, that is why it was rational of her to play Red-True, and it would have been irrational to play Blue-True. So she believes both that something is the case, and that it might not be the case. This seems irrational, but Lockeans insist that it is perfectly consistent with her being a model of rationality.

Back in subsection 2.3.1 I argued that this kind of thing would be a problem for any kind of orthodox theory. And in some sense all I'm doing here is noting that the Lockean really is a kind of orthodox theorist. But the argument that the Lockean is committed to the rationality of Moore-paradoxical claims doesn't rely on those earlier arguments; it's a direct consequence of their view applied to simple cases like Marie and Chamira.

6.4.4 Closure and the Lockean Theory

The Lockean theory makes an implausible prediction about conjunction.[8] It says that someone can believe two conjuncts, yet actively refuse to believe the conjunction. Here is how Stalnaker puts the point.

> Reasoning in this way from accepted premises to their deductive consequences (*P*, also *Q*, therefore *R*) does seem perfectly straightforward. Someone may object to one of the premises, or to the validity of the argument, but one could not intelligibly agree that the premises are each acceptable and the argument valid, while objecting to the acceptability of the conclusion. (Stalnaker, 1984, 92)

If believing that p just means having a credence in p above the threshold, then this will happen. Indeed, given some very weak assumptions about the world, it implies that there are plenty of quadruples $\langle S, A, B, A \wedge B\rangle$ such that

- S is a rational agent.
- A, B and $A \wedge B$ are propositions.
- S believes A and believes B.
- S does not believe $A \wedge B$.
- S knows that she has all these states, and consciously reflectively endorses them.

Now one might think, indeed I do think, that such quadruples do not exist at all. But set that objection aside. If the Lockean is correct, these quadruples should be everywhere. That's because for any $t \in (0, 1)$ you care to pick, quadruples of the form $\langle S, C, D, C \wedge D\rangle$ are very very common.

- S is a rational agent.
- C, D and $C \wedge D$ are propositions.
- S's credence in C is greater than t, and her credence in D is greater than t.
- S's credence in $C \wedge D$ is less than t.
- S knows that she has all these states, and reflectively endorses them.

The best arguments for the existence of quadruples $\langle S, A, B, A \wedge B\rangle$ are non-constructive existence proofs. David Christensen (2005) for instance,

[8] This subsection draws on material from my (2014).

argues from the existence of the preface paradox to the existence of these quadruples. I will come back to that argument in chapter 10. But what I want to stress here is that even if these existence proofs work, they don't really prove what the Lockean needs. They don't show that quadruples satisfying the constraints we associated with $\langle S, A, B, A \wedge B \rangle$ are just as common as quadruples satisfying the constraints we associated with $\langle S, C, D, C \wedge D \rangle$, for any t. But if the Lockean were correct, they should be exactly as common.

6.5 Belief as Probability One

6.6 Solving the Challenges

It's not fair to criticise other theories for their inability to meet a challenge that one's own theory cannot meet. So I'll end this chapter by noting that the six problems I've raised so far for Lockeans are not problems for my interest-relative theory of (rational) belief. I've already discussed the points about correctness in subsection 3.6.1, and about closure in chapter @(knowledge), and there isn't much to be added. But it's worth saying a few words about the other four problems.

6.6.1 Coins

I say that a necessary condition of believing that p is a disposition to take p for granted. The rational person will prefer betting on logically weaker rather than logically stronger propositions in the coin case, so they will not take the logically stronger ones for granted. If they did take them for granted, they would be indifferent between the bets. So they will not believe that one of the coin flips after the second will land heads, or even that one of the coin flips after the first will land heads. And that's the right result. The rational person should assign those propositions probability one, but not believe them.

6.6.2 Games

In the up-down game, if the rational person believed that the other player would play up, they would be indifferent between up and down. But it's irrational to be indifferent between those options, so they wouldn't have the belief. They will think the probability that the other person will play

up is one - what else could it be? But they will not believe it on pain of incoherence.

6.6.3 Arbitrariness

According to IRT, the difference between belief and non-belief is the difference between willingness and unwillingness to take something as given in inquiry. This is far from an arbitrary difference. And it is a difference that supports law-like generalisations. If someone believes that p, and believes that given p, A is better than B, they will prefer A to B. This isn't a universal truth; people make mistakes. But nor is it merely a statistical generalisation. Counterexamples are things to be explained, while instances are explained by the underlying pattern.

6.6.4 Moore

In many ways the guiding aim of this project was to avoid this kind of Moore paradoxicality. So it shouldn't be a surprise that we avoid it here. If someone shouldn't do something because p might be false, that's conclusive evidence that they don't know that p. And it's conclusive evidence that either they don't rationally believe p, or they are making some very serious mistake in their reasoning. And in the latter case, the reason they are making a mistake is not that p might be false, but that they have a seriously mistaken belief about the kind of choice they are facing. So we can never say that someone knows, or rationally believes, p, but their choice is irrational because p might be false.

Chapter 7

Hard Choices

The odds-based version of IRT is immune to a number of challenges that face stakes-based versions. But it has some of its own distinctive challenges. The most pressing of these concern choices between indiscriminable, or nearly indiscriminable, options.

Chapter 8

Stakes

This chapter discusses a lot of objections to interest-relative theories in epistemology that turn out to be primarily objections to stakes based versions of IRT. In most of the cases to be discussed, I think the objections are fairly good objections to that particular kind of IRT. But what many of the objectors fail to note is that this is just one kind of IRT, and an odds-based version is immune to these objections.

Chapter 9

Facing the Changes

Interest-relative theories are often accused of having intuitively implausible implications about when someone changes between knowing and not knowing something. And it certainly does have implications in this respect that are surprising if you have a certain kind of epistemological training. But given everything we have learned in the last 60 years about the kinds of thing knowledge is sensitive to, we shouldn't be surprised that knowledge is also sensitive to interests.

Chapter 10

The Preface Paradox

10.1 Solving the Paradox

In section ??, I argued against the Lockean thesis that full belief just is degree of belief above a threshold. And in particular I stressed in subsection 6.4.4 that the Lockean had a problem with closure. They were committed to the view that one could, with perfect rationality and perfect awareness, believe that p, believe that q, and decline to believe that $p \wedge q$. This seems absurd. It means that if they are answering sincerely, we can have the following conversation.

> Me: Is p true? Lockean: Yes. Me: Is q true? Lockean: Yes. Me: So p and q are both true? Lockean: Yes. Me: So $p \wedge q$ is true? Lockean: Well, not so fast, we can't be sure about that.

And that doesn't seem like a particularly good place to have ended up. But there is one prominent reason to end up there. David Christensen (2005) argues that reflection on the preface paradox shows that we need to have some kind of restriction on and-introduction. And the best restriction, he argues, will end up being one that makes the Lockean position look plausible.

Note that I'm here using 'and-introduction' as the name of a principle on rational inference. No one is denying that it is a good rule of implication. More precisely, I'm using it as the name of the following principle.

And-introduction If someone rationally believes p, and rationally believes q, and the question of whether $p \wedge q$ is true is a live one,

> then they are rationally entitled to believe $p \wedge q$, and if they decline to believe $p \wedge q$, they are not entitled to keep believing both p and q.

Here is Christensen's version of the preface paradox, which is meant to motivate a restriction on principles like and-introduction that link inference too tightly to logical implication.

> We are to suppose that an apparently rational person has written a long non-fiction book—say, on history. The body of the book, as is typical, contains a large number of assertions. The author is highly confident in each of these assertions; moreover, she has no hesitation in making them unqualifiedly, and would describe herself (and be described by others) as believing each of the book's many claims. But she knows enough about the difficulties of historical scholarship to realize that it is almost inevitable that at least a few of the claims she makes in the book are mistaken. She modestly acknowledges this in her preface, by saying that she believes the book will be found to contain some errors, and she graciously invites those who discover the errors to set her straight. (Christensen, 2005, 33-4)

Christensen thinks such an author might be rational in every one of her beliefs, even though these are all inconsistent. And he notes this is a quite ordinary belief. It's not like we have to go to fake barn country to find a counterexample to and-introduction. But it seems to me that we need two quite strong idealisations in order to get a real counterexample here.

The first of these is discussed by Ishani Maitra (2010), and is briefly mentioned by Christensen in setting out the problem. We only have a counterexample to and-introduction if the author believes every thing she writes in her book. Indeed, we only have a counterexample if she rationally believes every one of them. But we'll assume a rational author who only believes what she is rational to believe.

This seems unlikely. An author of a historical book is like a detective who, when asked to put forward her best guess about what explains the evidence, says "If I had to guess, I'd say …" and then launches into spelling out her hypothesis. It seems clear that she need not *believe* the truth of her hypothesis. If she did that, she could not later learn it was true, because you can't learn the truth of something you already believe. And she wouldn't put any effort into investigating alternative suspects. But she can

come to learn her hypothesis was true, and it would be rational to investigate other suspects. As Maitra suggests, we should understand scholarly assertions as being governed by the same kind of rules that govern detectives making the kind of speech being contemplated here. And those rules don't require that the speaker believe the things they say without qualification. The picture is that the little prelude the detective explicitly says is implicit in all scholarly work.

Christensen makes two objections to the suggestions that the author doesn't believe what she says. First, he notes that the author doesn't qualify their assertions. But neither does our detective qualify most individual sentences. Second, he notes that most people would describe our author as believing her assertions. But it is also natural to describe our detective as believing the things she says in her speech. It's natural to say things like "She thinks it was the butler, with the lead pipe, in the hallm" in reporting her hypothesis.

Alternatively, we might try to use an argument by Timothy Williamson (2000) to argue that speakers must believe what they say. He notes that if if speakers don't believe whatthey say, we won't have an explanation of why Moore paradoxical sentences like "The butler did it, but I don't believe the butler did it," are always defective. But note that our detective, who is explicitly asked for a guess, still sounds like she's making a mistake if she says "The butler did it, but I don't believe the butler did it." The detective can't even say that in setting out what is explicitly a hypothesis. So whatever explains why we can't say this in ordinary speech, it can't be that we are required to say what we believe.

It is plausible that for some kinds of books, the author should only say things they believe. This is probably true for travel guides, for example. Interestingly, casual observation suggests that authors of such books are much less likely to write modest prefaces. This makes some sense if those books can only include statements their authors believe, and the authors believe the conjunctions of what they believe.

The second idealisation is stressed by Simon Evnine (1999). The following situation does not involve Pedro believing anything inconsistent.

- Pedro believes that what Manny just said, whatever it was, is false.
- Manny just said that the stands at Fenway Park are green.
- Pedro believes that the stands at Fenway Park are green.

If we read the first claim de dicto, that Pedro believes that Manny just said something false, then there is no inconsistency. (Unless Pedro also

believes that what Manny just said was that the stands in Fenway Park are green.) But if we read it de re, that the thing Manny just said is one of the things Pedro believes to be false, then the situation does involve Pedro being inconsistent.

The same is true when the author believes that one of the things she says in her book is mistaken. If we understand what she says de dicto, there is no contradiction in her beliefs. It has to be understood de re before we get a logical problem. And the fact is that most authors do not have de re attitudes towards the claims made in their book. Most authors don't even remember everything that's in their books. (I'm not sure I remember how this chapter started, let alone this book.) Some may argue that authors don't even have the capacity to consider a proposition as long and complicated as the conjunction of all the claims in their book. Christensen considers this objection, but says it isn't a serious problem.

> It is undoubtedly true that ordinary humans cannot entertain book-length conjunctions. But surely, agents who do not share this fairly *superficial* limitation are easily conceived. And it seems just as wrong to say of such agents that they are rationally required to believe in the inerrancy of the books they write. (38: my emphasis)

We should be suspicious of the intuition that Christensen is relying on here. He argues idealising away from author's forgetfulness about what they've written shouldn't change our intuitions about the case. But the preface paradox gets a lot of its (apparent) force from intuitions about what attitude we should have towards real books. Once we make it clear that the real life cases are not relevant to the paradox, the intuitions become rather murky. And the idea that there are failures of and-introduction that arise in everyday cases, that aren't like weird fake barn cases, has falled away a little at this point.

Ultimately, I think the first idealisation is more important. We believers in and-introduction don't think that authors should believe their books are inerrant. Rather, following a position that stretches back at least to Stalnaker (1984), they shouldn't fully believe each individual statement if they don't believe the conjunction. They don't have to have any particular quantified attitude to each claim in the book. They can simply think that each claim is probably true, and decline to associate this attitude with any numerical credence.

Proponents of the preface paradox know that this is a possible response. The standard response is that it is impractical. Here is Christensen on this point.

> It is clear that our everyday binary way of talking about beliefs has immense practical advantages over a system which insisted on some more fine-grained reporting of degrees of confidence … At a minimum, talking about people as believing, disbelieving, or withholding belief has at least as much point as do many of the imprecise ways we have of talking about things that can be described more precisely. (96)

Richard Foley makes a similar point.

> There are deep reasons for wanting an epistemology of beliefs, reasons that epistemologies of degrees of belief by their very nature cannot possibly accommodate. (Foley, 1993, 170, my emphasis)

It's easy to make too much of this point. It's a lot easier to triage propositions into TRUE, FALSE and NOT SURE and work with those categories than it is to work assign precise numerical probabilities to each proposition. But these are not the only options. Foley's discussion subsequent to the above quote sometimes suggests they are, especially when he contrasts the triage with "indicat

ing

as accurately as I can my degree of confidence in each assertion that I defend." (171)

But really it isn't much harder to add two more categories, PROBABLY TRUE and PROBABLY FALSE to those three, and work with that five-way division rather than a three-way division. It's not clear that humans as they are actually constructed have a strong preference for the three-way over the five-way division, and even if they do, I'm not sure in what sense this is a 'deep' fact about them.

Once we have the five-way division, it is clear what authors should do if they want to respect and-introduction. For any conjunction that they don't believe (i.e. classify as true), they should not believe one of the conjuncts. But of course they can classify every conjunct as probably true, even if they think the conjunction is false, or even certainly false. Still,

might it not be considered something of an idealisation to say rational authors must make this five-way distinction amongst propositions they consider? Yes, but it's no more of an idealisation than we need to set up the preface paradox in the first place. To use the preface paradox to find an example of someone who reasonably violates closure, we need to insist on the following three constraints.

1. They are part of a research community where only asserting propositions you believe is compatible with active scholarship;
2. They know exactly what is in their book, so they are able to believe that one of the propositions in the book is mistaken, where this is understood de re; but
3. They are unable to effectively function if they have to effect a five-way, rather than a three-way, division amongst the propositions they consider.

Put more graphically, to motivate the preface paradox we have to think that our inability to have de re thoughts about the contents of books is a "superficial constraint", but our preference for working with a three-way rather than a five-way division is a "deep" fact about our cognitive system. Maybe each of these attitudes could be plausible taken on its own (though I'm sceptical of that) but the conjunction is very hard to motivate.

Even if someone who satisfied precisely these idealisations was possible, something that isn't at all obvious to me, it isn't clear why the norms applicable to them have any relevance to us. That is, it might be that for someone who satisfied 1, 2 and 3, then violating and-introduction would be a way to make the best of a bad situation. But it wouldn't follow that violating and-introduction is rationally permissible. It might be that this is the least bad way to deal with the constraints that have been imposed.

10.2 Too Little Closure?

At this point, one might worry that I've argued for a conclusion that is stronger than what I wanted to defend. While my version of IRT endorses and-introduction as stated, it does not endorse a version with some of the restrictions removed. In particular, it doesn't endorse a version of the view that says whenever a rational person believes p and believes q, they also believe $p \wedge q$. And I rather doubt one could endorse that, while holding on to anything like the picture I'm presenting here.

So I'll end by defending these restrictions. Let's start by looking at a very important argument for (something like) and-introduction, taken from Stalnaker's *Inquiry*.

> Reasoning in this way from accepted premises to their deductive consequences (*P*, also *Q*, therefore *R*) does seem perfectly straightforward. Someone may object to one of the premises, or to the validity of the argument, but one could not intelligibly agree that the premises are each acceptable and the argument valid, while objecting to the acceptability of the conclusion. (Stalnaker, 1984, 92)

Stalnaker's wording here is typically careful. The relevant question isn't whether we can accept *p*, accept *q*, accept *p* and *q* entail *r*, and reject *r*. As Christensen (2005, Ch. 4) notes, this is impossible even on the Lockean view, as long as the threshold for belief is above 2/3. The real question is whether we can accept *p*, accept *q*, accept *p* and *q* entail *r*, and fail to accept *r*. And this is always a live possibility on any version of the Lockean view.

But it's important to note how active the verbs in Stalnaker's description are. When faced with a valid argument we have to *object* to one of the premises, or the validity of the argument. What we can't do is *agree* to the premises and the validity of the argument, while *objecting* to the conclusion. I agree. If we are really agreeing to some propositions, and objecting to others, then all those propositions are live in the sense relevant to and-introduction. And in that case believing the conjunction of whatever is believed is mandatory. This doesn't tell us what we have to do if we haven't previously made the propositions salient in the first place.

Where I've ended up is very similar to a position that Gilbert Harman endorses in *Change in View*. There Harman endorses the following principle. (At least he endorses it as true – he doesn't seem to think it is particularly explanatory because it is a special case of a more general interesting principle.)

> One has reason to believe *P* if one *recognizes* that *P* is logically implied by one's view. (Harman, 1986, 17)

This is, like the passage from Stalnaker I just quoted, both correct and careful. Notably, it does not say that this reason is a conclusive reason. After all, one might change one's view rather than accept the implication. But one does have reason to believe *p* in such a situation.

My main objection to those who use the preface paradox to argue against and-introduction is that they give us a mistaken picture of what we have to do epistemically. When one has inconsistent beliefs, or one doesn't believe some consequence of one's beliefs, that is something one has a reason to deal with at some stage. It is something that is to do for one; i.e., it should be on one's 'to-do' list.

When we say that we have things to do, we don't mean that we have to do them *right now*, or instead of everything else. My current list of things to do includes cleaning my office. Yet I'm working on this book and, given the relative importance of a completed book and a clean office, rightly so.

We can have the job of cleaning up our epistemic house as something to do while recognising that we can quite rightly do other things first. But it's a serious mistake to infer from the permissibility of doing other things that cleaning up our epistemic house (or our office) isn't something to be done. The office won't clean itself after all, and eventually this becomes a problem.

There is a possible complication when it comes to tasks that are very low priority. Imagine someone with an attic that is both somewhat messy, and also rarely used because it is so impractical. It is, in a sense, to be cleaned. At least, it could be cleaner. But there are no imaginable circumstances under which something else wouldn't be higher priority. Given that, should we really leave *clean the attic* on the list of things to be done? Similarly, there might be implications of one's beliefs that one hasn't deduced that it couldn't possibly be worth my time to figure out. Are they things to be done? I think it's worthwhile recording them as such, because otherwise we might miss opportunities to deal with them in the process of doing something else. I don't need to put off anything else in order to clean the attic, but if I'm up there for independent reasons I should bring down some of the junk. Similarly, I don't need to follow through implications mostly irrelevant to my interests, but if those propositions come up for independent reasons, I should deal with the fact that some things I believe imply something I don't believe. Having it be the case that all implications from things we believe to things we don't believe constitute jobs to do (possibly in the loose sense that cleaning my attic is something to do) has the right implications for what epistemic duties we do and don't have.

While waxing metaphorical, it seems time to pull out a rather helpful Rylean metaphor. It's from the discussion in Ryle (1949) of what we'd now call the inference/implication distinction. (This is a large theme of

chapter 9, see particularly pages 292-309.) Ryle's point in these passages, as it frequently is throughout the book, is to stress that minds are fundamentally active, and the activity of a mind cannot be easily recovered from its end state. Although Ryle doesn't use this language, his point is that we shouldn't confuse the difficult activity of drawing inferences with the smoothness and precision of a logical implication. The language Ryle does use is more picturesque. He compares the easy work a farmer does when sauntering down a path from the hard work he did when building the path. A good argument, in philosophy or mathematics or elsewhere, is like a well made path that permits sauntering from the start to finish without undue strain. But from that it doesn't follow that the task of coming up with that argument, of building that path in Ryle's metaphor, was easy work. The easiest paths to walk are often the hardest to build. Path-building, smoothing out our beliefs so they are consistent and closed under implication, is hard work, even when the finished results look clean and straightforward. Its work that we shouldn't do unless we need to. But making sure our beliefs are closed under entailment even with respect to irrelevant propositions is suspiciously like the activity of buildings paths between points without first checking you need to walk between them.

So it's fine to believe p, believe q, and fail to believe $p \wedge q$, as long as it isn't the case that all three of those propositions are relevant to one's intersts at any one time. There is a tension there, and it is a tension that is to be relieved eventually. But the time to relieve it might never arise, and when it does, there are a lot of ways to deal with the situation.

Chapter 11

Conclusion

I'm right about everything and everyone else is wrong.

Bibliography

Anderson, Charity and Hawthorne, John. 2019a. "Knowledge, Practical Adequacy, and Stakes." *Oxford Studies in Epistemology* 6: 234-257.

—. 2019b. "Pragmatic Encroachment and Closure." In Kim and McGrath (2019), 107-115.

Armour-Garb, B. 2011. "Contextualism Without Pragmatic Encroachment." *Analysis* 71: 667-676, doi:10.1093/analys/anr083.

Aumann, Robert J. 1999. "Interactive Epistemology I: Knowledge." *International Journal of Game Theory* 28: 263-300, doi:10.1007/s001820050111.

Bhatt, Rajesh. 1999. *Covert Modality in Non-finite Contexts*. Ph.D. thesis, University of Pennsylvania.

Binmore, Ken. 2007. *Playing for Real: A Text on Game Theory*. Oxford: Oxford University Press.

Boyd, Kenneth. 2016. "Pragmatic Encroachment and Epistemically Responsible Action." *Synthese* 193: 2721-2745, doi:10.1007/s11229-015-0878-y.

Brown, Jessica. 2008. "Subject-Sensitive Invariantism and the Knowledge Norm for Practical Reasoning." *Noûs* 42: 167-189, doi:10.1111/j.1468-0068.2008.00677.x.

Carlsson, Hans and van Damme, Eric. 1993. "Global Games and Equilibrium Selection." *Econometrica* 61: 989-1018, doi:10.2307/2951491.

Christensen, David. 2005. *Putting Logic in Its Place*. Oxford: Oxford University Press.

Clark, Christopher. 2012. *The Sleepwalkers: How Europe Went to War in 1914*. New York: Harper Collins.

Cohen, Stewart. 2004. "Knowledge, assertion, and practical reasoning." *Philosophical Issues* 14: 482-491, doi:10.1111/j.1533-6077.2004.00040.x.

Cross, Charles and Roelofsen, Floris. 2018. "Questions." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2018 edition.

DeRose, Keith. 2002. "Assertion, Knowledge and Context." *Philosophical Review* 111: 167-203, doi:10.2307/3182618.

Dixit, Avinash K. and Skeath, Susan. 2004. *Games of Strategy*. New York: W. W. Norton & Company, second edition.

Evnine, Simon. 1999. "Believing Conjunctions." *Synthese* 118: 201-227, doi:10.1023/A:1005114419965.

Fantl, Jeremy and McGrath, Matthew. 2002. "Evidence, Pragmatics, and Justification." *Philosophical Review* 111: 67-94, doi:10.2307/3182570.

—. 2009. *Knowledge in an Uncertain World*. Oxford: Oxford University Press.

Foley, Richard. 1993. *Working Without a Net*. Oxford: Oxford University Press.

Ganson, Dorit. 2008. "Evidentialism and Pragmatic Constraints on Outright Belief." *Philosophical Studies* 139: 441-458, doi:10.1007/s11098-007-9133-9.

—. 2019. "Great Expectations: Belief and the Case for Pragmatic Encroachment." In Kim and McGrath (2019).

Gendler, Tamar Szabó and Hawthorne, John. 2005. "The Real Guide to Fake Barns: A Catalogue of Gifts for Your Epistemic Enemies." *Philosophical Studies* 124: 331-352, doi:10.1007/s11098-005-7779-8.

Gettier, Edmund L. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23: 121-123, doi:10.2307/3326922.

Gillies, Anthony S. 2010. "Iffiness." *Semantics and Pragmatics* 3: 1-42, doi:10.3765/sp.3.4.

Harman, Gilbert. 1973. *Thought*. Princeton: Princeton University Press.

—. 1986. *Change in View*. Cambridge, MA: Bradford.

Hawthorne, John. 2004. *Knowledge and Lotteries*. Oxford: Oxford University Press.

—. 2005. "Knowledge and Evidence." *Philosophy and Phenomenological Research* 70: 452-458, doi:10.1111/j.1933-1592.2005.tb00540.x.

Humberstone, I. L. 1981. "From Worlds to Possibilities." *Journal of Philosophical Logic* 10: 313-339, doi:10.1007/BF00293423.

Hunter, Daniel. 1996. "On the Relation Between Categorical and Probabilistic Belief." *Noûs* 30: 75-98, doi:10.2307/2216304.

Kelly, Thomas. 2010. "Peer disagreement and higher order evidence." In Ted Warfield and Richard Feldman (eds.), *Disagreement*, 111-174. Oxford: Oxford University Press.

Kim, Brian and McGrath, Matthew (eds.). 2019. *Pragmatic Encroachment in Epistemology*. New York: Routledge.

Kohlberg, Elon and Mertens, Jean-Francois. 1986. "On the Strategic Stability of Equilibria." *Econometrica* 54: 1003-1037, doi:10.2307/1912320.

Kratzer, Angelika. 2012. *Modals and Conditionals*. Oxford: Oxford University Press.

Lasonen-Aarnio, Maria. 2010. "Unreasonable Knowledge." *Philosophical Perspectives* 24: 1-21, doi:10.1111/j.1520-8583.2010.00183.x.

—. 2014. "Higher-Order Evidence and the Limits of Defeat." *Philosophy and Phenomenological Research* 88: 314-345, doi:10.1111/phpr.12090.

Lee, Matthew. 2017a. "On the Arbitrariness Objection to the Threshold View." *Dialogue* 56: 143-158, doi:10.1017/S0012217317000154.

Lee, Matthew Brandon. 2017b. "Credence and Correctness: In Defense of Credal Reductivism." *Philosophical Papers* 46: 273-296, doi:10.1080/05568641.2017.1364142.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

—. 1976. "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review* 85: 297-315, doi:10.2307/2184045. Reprinted in *Philosophical Papers*, Volume II, pp. 133-152.

—. 1986. "Probabilities of Conditionals and Conditional Probabilities II." *Philosophical Review* 95: 581-589, doi:10.2307/2185051. Reprinted in *Papers in Philosophical Logic*, pp. 57-65.

—. 1988. "Desire as Belief." *Mind* 97: 323-332, doi:10.1093/mind/xcvii.387.323.

—. 1996. "Desire as Belief II." *Mind* 105: 303-313, doi:10.1093/mind/105.418.303.

—. 2004. "Causation as Influence." In John Collins, Ned Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*, 75-106. Cambridge: MIT Press.

Maitra, Ishani. 2010. "Assertion, Norms and Games." In Jessica Brown and Herman Cappelen (eds.), *Assertion: New Philosophical Essays*, 277-296. Oxford: Oxford University Press.

Maitra, Ishani and Weatherson, Brian. 2010. "Assertion, Knowledge and Action." *Philosophical Studies* 149: 99-118, doi:10.1007/s11098-010-9542-z.

McGrath, Matthew and Kim, Brian. 2019. "Introduction." In Kim and McGrath (2019), 1-9.

Nagel, Jennifer. 2008. "Knowledge ascriptions and the psychological consequences of changing stakes." *Australasian Journal of Philosophy* 86: 279-294, doi:10.1080/00048400801886397.

—. 2010. "Epistemic Anxiety and Adaptive Invariantism." *Philosophical Perspectives* 24: 407-435, doi:10.1111/j.1520-8583.2010.00198.x.

—. 2013. "Motivating Williamson's Model Gettier Cases." *Inquiry* 56: 54-62, doi:10.1080/0020174X.2013.775014.

—. 2014. *Knowledge: A Very Short Introduction*. Oxford: Oxford University Press.

North, Jill. 2010. "An empirical approach to symmetry and probability." *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics* 41: 27-40, doi:10.1016/j.shpsb.2009.08.008.

Nozick, Robert. 1981. *Philosophical Explorations*. Cambridge, MA: Harvard University Press.

Ramsey, Frank. 1990. "General Propositions and Causality." In D. H. Mellor (ed.), *Philosophical Papers*, 145-163. Cambridge: Cambridge University Press.

Richter, Reed. 1984. "Rationality Revisited." *Australasian Journal of Philosophy* 62: 393-404, doi:10.1080/00048408412341601.

Roberts, Craige. 2012. "Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics." *Semantics and Pragmatics* 5: 1-69, doi:10.3765/sp.5.6.

Ross, Jacob and Schroeder, Mark. 2014. "Belief, Credence, and Pragmatic Encroachment." *Philosophy and Phenomenological Research* 88: 259-288, doi:10.1111/j.1933-1592.2011.00552.x.

Rousseau, Jean-Jacques. 1913. *Social Contract & Discourses*. New York: J. M. Dent & Sons. Translated by G. D. H. Cole.

Ryle, Gilbert. 1949. *The Concept of Mind*. New York: Barnes and Noble.

Schoenfield, Miriam. 2013. "Permission to Believe: Why Permissivism Is True and What It Tells Us About Irrelevant Influences on Belief." *Noûs* 47: 193-218, doi:10.1111/nous.12006.

Schwitzgebel, Eric. 2008. "The Unreliability of Naive Introspection." *Philosophical Review* 117: 245-273, doi:10.1215/00318108-2007-037.

Sperber, Dan, Clément, Fabrice, Heintz, Christophe, Mascaro, Olivier, Mercier, Hugo, Origgi, Gloria, and Wilson, Deirdre. 2010. "Epistemic Vigilance." *Mind and Language* 25: 359-393, doi:10.1111/j.1468-0017.2010.01394.x.

Stalnaker, Robert. 1984. *Inquiry*. Cambridge, MA: MIT Press.

—. 1994. "On the evaluation of solution concepts." *Theory and Decision* 37: 49-73, doi:10.1007/BF01079205.

—. 1996. "Knowledge, Belief and Counterfactual Reasoning in Games." *Economics and Philosophy* 12: 133-163, doi:10.1017/S0266267100004132.

—. 1998. "Belief revision in games: forward and backward induction." *Mathematical Social Sciences* 36: 31-56, doi:10.1016/S0165-4896(98)00007-9.

—. 1999. "Extensive and strategic forms: Games and models for games." *Research in Economics* 53: 293-319, doi:10.1006/reec.1999.0200.

Stanley, Jason. 2005. *Knowledge and Practical Interests*. Oxford University Press.

—. 2011. *Know How*. Oxford: Oxford University Press.

Unger, Peter. 1975. *Ignorance: A Case for Scepticism*. Oxford: Oxford University Press.

Weatherson, Brian. 2005a. "Can We Do Without Pragmatic Encroachment?" *Philosophical Perspectives* 19: 417-443, doi:10.1111/j.1520-8583.2005.00068.x.

—. 2005b. "True, Truer, Truest." *Philosophical Studies* 123: 47-70, doi:10.1007/s11098-004-5218-x.

—. 2012. "Knowledge, Bets and Interests." In Jessica Brown and Mikkel Gerken (eds.), *Knowledge Ascriptions*, 75-103. Oxford: Oxford University Press.

—. 2014. "Games, Beliefs and Credences." *Philosophy and Phenomenological Research* 92: 209-236, doi:10.1111/phpr.12088.

—. 2016. "Games, Beliefs and Credences." *Philosophy and Phenomenological Research* 92: 209-236, doi:10.1111/phpr.12088.

—. 2019. *Normative Externalism*. Oxford: Oxford University Press.

Weisberg, Jonathan. 2013. "Knowledge in Action." *Philosophers' Imprint* 13: 1-23, doi:2027/spo.3521354.0013.022.

—. 2020. "Belief in Psyontology." *Philosophers' Imprint* xx-xx.

White, Roger. 2005. "Epistemic permissiveness." *Philosophical Perspectives* 19: 445-459, doi:10.1111/j.1520-8583.2005.00069.x.

Williamson, Timothy. 1994. *Vagueness*. New York: Routledge.

—. 2000. *Knowledge and its Limits*. Oxford University Press.

—. 2007. "How probable is an infinite sequence of heads?" *Analysis* 67: 173-180, doi:10.1111/j.1467-8284.2007.00671.x.

—. 2013. "Gettier Cases in Epistemic Logic." *Inquiry* 56: 1-14, doi:10.1080/0020174X.2013.775010.

Zagzebski, Linda. 1994. "The Inescapability of Gettier Problems." *The Philosophical Quarterly* 44: 65-73, doi:10.2307/2220147.

Zweber, Adam. 2016. "Fallibilism, Closure, and Pragmatic Encroachment." *Philosophical Studies* 173: 2745-2757, doi:10.1007/s11098-016-0631-5.