Belief, Knowledge and Interests

Brian Weatherson

2013

About this Document

Draft towards a book on interest-relativity, belief and knowledge.

Introduction

What the view is:

- An interest-relative theory of belief
- An interest-relative theory of knowledge
- Used to hope that the second was solely derived from the first; now not so clear

What's distinctive about the view:

- Invariantist
- Grounded in a theory of belief
- Integrated with game theory and decision theory; supports, and supported by, an epistemic interpretation of some problematic parts of game theory and decision theory.

Contents

Chapter 1

Defining the Target

Chapter 2

Belief and Interests

2.1 *Belief and Degree of Belief*

Traditional epistemology deals with beliefs and their justification. Bayesian epistemology deals with degrees of belief and their justification. In some sense they are both talking about the same thing, namely epistemic justification. Two questions naturally arise. Do we really have two subject matters here (degrees of belief and belief *tout court*) or two descriptions of the one subject matter? If just one subject matter, what relationship is there between the two modes of description of this subject matter?

The answer to the first question is I think rather easy. There is no evidence to believe that the mind contains two representational systems, one to represent things as being probable or improbable and the other to represent things as being true or false. The mind probably does contain a vast plurality of representational systems, but they don't divide up the doxastic duties this way. If there are distinct visual and auditory representational systems, they don't divide up duties between degrees of belief and belief *tout court*, for example. If there were two distinct systems, then we should imagine that they could vary independently, at least as much as is allowed by constitutive rationality. But such variation is hard to fathom. So I'll infer that the one representational system accounts for our credences and our categorical beliefs. (It follows from this that the question Bovens and Hawthorne (1999) ask, namely what beliefs *should* an agent have given her degrees of belief, doesn't have a non-trivial answer. If fixing the degrees of belief in an environment fixes all her doxastic attitudes, as I think it does, then there is no further question of what she should believe given these are her degrees of belief.)

The second question is much harder. It is tempting to say that S believes that p iff S's credence in p is greater than some salient number r, where r is made salient either by the context of belief ascription, or the context that S is in. I'm following Mark Kaplan (1996) in calling this the threshold view. There are two well-known problems with the threshold view, both of which seem fatal to me.

As Robert Stalnaker (1984, 91) emphasised, any number r is bound to seem arbitrary. Unless these numbers are made salient by the environment, there is no special difference between believing p to degree 0.9786 and believing it to degree 0.9875. But if r 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) is to say that the boundary is vague. But it's not clear how this helps. On an epistemic theory of vagueness, there is still a number such that degrees of belief above that count, and degrees below that do not, and any such number is bound to seem unimportant. On supervaluational theories, the same is true. There won't be a *determinate* number, to be sure, but there will a number, and that seems false. My preferred degree of belief theory of vagueness, as set out in Weatherson (2005b) has the same consequence. Hunter defends a version of the threshold view combined with a theory of vagueness based around fuzzy logic, which seems to be the only theory that could avoid the arbitrariness objection. But as Williamson (1994) showed, there are deep and probably insurmountable difficulties with that position. So I think the vagueness response to the arbitrariness objection is (a) the only prima facie plausible response and (b) unsuccessful.

The second problem concerns conjunction. It is also set out clearly by Stalnaker.

> 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 categorical belief is having a credence above the threshold, then one can coherently do exactly this. Let x be a number between r and than $r^{1/2}$, such that for an atom of type U has probability x of decaying within a time t, for some t and U. Assume our agent knows this fact, and is faced with two (isolated) atoms of U. Let p be that the first decays within t, and q be that the second decays within t. She should, given her evidence, believe p to degree x, q to degree x, and $p \wedge q$ to degree x^2. If she believed $p \wedge q$ to a degree greater than r, she'd have to either have credences that were not supported by her evidence, or credences that were incoherent. (Or, most likely, both.) So this theory violates the platitude. This is a well-known argument, so there are many responses to it, most of them involving something like appeal to the preface paradox. I'll argue in section 4 that the preface paradox doesn't in fact offer the threshold view proponent much support here. But even before we get to there, we should note that the arbitrariness objection gives us sufficient reason to reject the threshold view.

A better move is to start with the functionalist idea that to believe that p is to treat p as true for the purposes of practical reasoning. To believe p is to have preferences that make sense, by your own lights, in a world where p is true. So, if you prefer A to B and believe that p, you prefer A to B given p. For reasons that will become apparent below, we'll work in this paper with a notion of preference where *conditional* preferences are primary.[1] So the core insight we'll work with is the following:

> If you prefer A to B given q, and you believe that p, then you prefer A to B given $p \wedge q$

The bold suggestion here is that if that is true for all the A, B and q that matter, then you believe p. Put formally, where $Bel(p)$ means that the agent believes that p, and $\mathrm{A} \geq_q \mathrm{B}$ means that the agent thinks A is at least as good as B given q, we have the following

1. $Bel(p) \leftrightarrow \forall \mathrm{A} \forall \mathrm{B} \forall q\ (\mathrm{A} \geq_q \mathrm{B} \leftrightarrow \mathrm{A} \geq_{p \wedge q} \mathrm{B})$

In words, an agent believes that p iff conditionalising on p doesn't change any conditional preferences over things that matter.[2] The left-to-right direction of this seems trivial, and the right-to-left direction seems to be a plausible way to operationalise the functionalist insight that belief is a functional state. There is some work to be done if (1) is to be interpreted as a truth though.

If we interpret the quantifiers in (1) as unrestricted, then we get the (false) conclusion that just about no one believes no contingent propositions. To prove this, consider a bet that wins iff the statue in front of me waves back at me due to random quantum effects when I wave at it. If I take the bet and win, I get to

[1] To say the agent prefers A to B given q is not to say that if the agent were to learn q, she would prefer A to B. It's rather to say that she prefers the state of the world where she does A and q is true to the state of the world where she does B and q is true. These two will come apart in cases where learning q changes the agent's preferences. We'll return to this issue below.

[2] This might seem *much* too simple, especially when compared to all the bells and whistles that functionalists usually put in their theories to (further) distinguish themselves from crude versions of behaviourism. The reason we don't need to include those complications here is that they will all be included in the analysis of *preference*. Indeed, the theory here is compatible with a thoroughly anti-functionalist treatment of preference. The claim is not that we can offer a functional analysis of belief in terms of non-mental concepts, just that we can offer a functionalist reduction of belief to other mental concepts. The threshold view is *also* such a reduction, but it is such a crude reduction that it doesn't obviously fall into any category.

live forever in paradise. If I take the bet and lose, I lose a penny. Letting A be that I take the bet, B be that I decline the bet, q be a known tautology (so my preferences given q are my preferences *tout court*) and p be that the statue does not wave back, we have that I prefer A to B, but not A to B given p. So by this standard I don't believe that p. This is false – right now I believe that statues won't wave back at me when I wave at them.

This seems like a problem. But the solution to it is not to give up on functionalism, but to insist on its pragmatic foundations. The quantifiers in (1) should be restricted, with the restrictions motivated pragmatically. What is crucial to the theory is to say what the restrictions on A and B are, and what the restrictions on q are. We'll deal with these in order.

For better or worse, I don't right now have the option taking that bet and hence spending eternity in paradise if the statue waves back at me. Taking or declining such unavailable bets are not open choices. For any option that is open to me, assuming that statues do not in fact wave does not change its utility. That's to say, I've already factored in the non-waving behaviour of statues into my decision-making calculus. That's to say, I believe statues don't wave.

An action A is a live option for the agent if it is really possible for the agent to perform A. An action A is a salient option if it is an option the agent takes seriously in deliberation. Most of the time gambling large sums of money on internet gambling sites over my phone is a live option, but not a salient option. I know this option is suboptimal, and I don't have to recompute every time whether I should do it. Whenever I'm making a decision, I don't have to add in to the list of choices *bet thousands of dollars on internet gambling sites*, and then rerule that out every time. I just don't consider that option, and properly so. If I have a propensity to daydream, then becoming the centrefielder for the Boston Red Sox might be a salient option to me, but it certainly isn't a live option. We'll say the two initial quantifiers range over the options that are live and salient options for the agent.

Note that we *don't* say that the quantifiers range over the options that are live and salient for the person making the belief ascription. That would lead us to a form of contextualism for which we have little evidence. We also don't say that an option becomes salient for the agent iff they *should* be considering it. At this stage we are just saying what the agent does believe, not what they should believe, so we don't have any clauses involving normative concepts.

Now we'll look at the restrictions on the quantifier over propositions. Say a proposition is *relevant* if the agent is disposed to take seriously the question of whether it is true (whether or not she is currently considering that question) and conditionalising on that proposition or its negation changes some of the agents *unconditional* preferences over live, salient options.[3] The first clause is designed to rule out wild hypotheses that the agent does not take at all seriously. If q is not such a proposition, if the agent is disposed to take it seriously, then it is relevant if there are live, salient A and B such that $A \geq_q B \leftrightarrow A \geq B$ is false. Say a proposition is *salient* if the agent is currently considering whether it is true. Finally, say a proposition is *active* relative to p iff it is a (possibly degenerate) conjunction of propositions such that each conjunct is either relevant or salient, and such that the conjunction is consistent with p. (By a degenerate conjunction I mean a conjunction with just one conjunct. The consistency requirement is there because it might be hard in some cases to make sense of preferences given inconsistencies.) Then the propositional quantifier in (1) ranges over active propositions.

We will expand and clarify this in the next section, but our current solution to the relationship between beliefs and degrees of belief is that degrees of belief determine an agent's preferences, and she believes that p iff the claim (1) about her preferences is true when the quantifiers over options are restricted to live, salient actions, and the quantifier over propositions is restricted to salient propositions. The simple view would be to say that the agent believes that p iff conditioning on p changes none of her preferences. The more complicated view here is that the agent believes that p iff conditioning on p changes none of her conditional preferences over live, salient options, where the conditions are also active relative to p.

2.2 *Impractical Propositions*

The theory sketched in the previous paragraph seems to me right in the vast majority of cases. It fits in well with a broadly functionalist view of the mind, and as we'll see it handles some otherwise difficult cases with aplomb. But it needs to be supplemented a little to handle beliefs about propositions that are practically irrelevant. I'll illustrate the problem, then note how I prefer to solve it.

I don't know what Julius Caeser had for breakfast the morning he crossed the Rubicon. But I think he would have had *some* breakfast. It is hard to be a good general without a good morning meal after all. Let p be the proposition that he had breakfast that morning. I believe p. But this makes remarkably little difference to my practical choices in most situations. True, I wouldn't have

[3] Conditionalising on the proposition *There are space aliens about to come down and kill all the people writing epistemology papers* will make me prefer to stop writing this paper, and perhaps grab some old metaphysics papers I could be working on. So that proposition satisfies the second clause of the definition of relevance. But it clearly doesn't satisfy the first clause. This part of the definition of relevance won't do much work until the discussion of agents with mistaken environmental beliefs in section 7.

written this paragraph as I did without this belief, but it is rare that I have to write about Caeser's dietary habits. In general whether p is true makes no practical difference to me. This makes it hard to give a pragmatic account of whether I believe that p. Let's apply (1) to see whether I really believe that p.

1. $\mathit{Bel}(p) \leftrightarrow \forall \mathrm{A} \forall \mathrm{B} \forall q\ (\mathrm{A} \geq_q \mathrm{B} \leftrightarrow \mathrm{A} \geq_{p \wedge q} \mathrm{B})$

Since p makes no practical difference to any choice I have to make, the right hand side is true. So the left hand side is true, as desired. The problem is that the right hand side of (2) is also true here.

2. $\mathit{Bel}(\neg p) \leftrightarrow \forall \mathrm{A} \forall \mathrm{B} \forall q\ (\mathrm{A} \geq_q \mathrm{B} \leftrightarrow \mathrm{A} \geq_{\neg p \wedge q} \mathrm{B})$

Adding the assumption that Caeser had no breakfast that morning doesn't change any of my practical choices either. So I now seem to *inconsistently* believe both p and $\neg p$. I have some inconsistent beliefs, I'm sure, but those aren't among them. We need to clarify what (1) claims.

To do so, I supplement the theory sketched in section 2 with the following principles.

- A proposition p is *eligible for belief* if it satisfies $\forall \mathrm{A} \forall \mathrm{B} \forall q\ (\mathrm{A} \geq_q \mathrm{B} \leftrightarrow \mathrm{A} \geq_{p \wedge q} \mathrm{B})$, where the first two quantifiers range over the open, salient actions in the sense described in section 2.
- For any proposition p, and any proposition q that is relevant or salient, among the actions that are (by stipulation!) open and salient with respect to p are *believing that p*, *believing that q*, *not believing that p* and *not believing that q*
- For any proposition, the subject prefers believing it to not believing it iff (a) it is eligible for belief and (b) the agents degree of belief in the proposition is greater than 1/2.
- The previous stipulation holds both unconditionally and conditional on p, for any p.
- The agent believes that p iff $\forall \mathrm{A} \forall \mathrm{B} \forall q\ (\mathrm{A} \geq_q \mathrm{B} \leftrightarrow \mathrm{A} \geq_{p \wedge q} \mathrm{B})$, where the first two quantifiers range over all actions that are either open and salient *tout court* (i.e. in the sense of section 2) or open and salient with respect to p (as described above).

This all looks moderately complicated, but I'll explain how it works in some detail as we go along. One simple consequence is that an agent only believes that p iff their degree of belief in p is greater than 1/2. Since my degree of belief in Caeser's foodless morning is not greater than 1/2, in fact it is considerably less, I don't believe $\neg p$. On the other hand, since my degree of belief in p is considerably greater than 1/2, I prefer to believe it than disbelieve it, so I believe it.

There are many possible objections to this position, which I'll address sequentially.

Objection: Even if I have a high degree of belief in p, I might prefer to not believe p because I think that belief in p is bad for some other reason. Perhaps, if p is a proposition about my brilliance, it might be immodest to believe that p.
Reply: Any of these kinds of considerations should be put into the credences. If it is immodest to believe that you are a great philosopher, it is equally immodest to believe to a high degree that you are a great philosopher.

Objection: Belief that p is not an action in the ordinary sense of the term.
Reply: True, which is why this is described as a supplement to the original theory, rather than just cashing out its consequences.

Objection: It is impossible to choose to believe or not believe something, so we shouldn't be applying these kinds of criteria.
Reply: I'm not as convinced of the impossibility of belief by choice as others are, but I won't push that for present purposes. Let's grant that beliefs are always involuntary. So these 'actions' aren't open actions in any interesting sense, and the theory is section 2 was really incomplete. As I said, this is a supplement to the theory in section 2.

This doesn't prevent us using principles of constitutive rationality, such as we prefer to believe p iff our credence in p is over 1/2. Indeed, on most occasions where we use constitutive rationality to infer that a person has some mental state, the mental state we attribute to them is one they could not fail to have. But functionalists are committed to constitutive rationality (Lewis, 1994). So my approach here is consistent with a broadly functionalist outlook.

Objection: This just looks like a roundabout way of stipulating that to believe that p, your degree of belief in p has to be greater than 1/2. Why not just add that as an extra clause than going through these little understood detours about preferences about beliefs?
Reply: There are three reasons for doing things this way rather than adding such a clause.

First, it's nice to have a systematic theory rather than a theory with an ad hoc clause like that.

Second, the effect of this constraint is much more than to restrict belief to propositions whose credence is greater than 1/2. Consider a case where p and q and their conjunction are all salient, p and q are probabilistically independent, and the agent's credence in each is 0.7. Assume also that p, q and $p \wedge q$ are completely irrelevant to any practical deliberation the agent must make. Then the criteria above imply that the agent does not believe that p or that q. The reason is that the agent's credence in $p \wedge q$ is 0.49, so she prefers to not believe $p \wedge q$. But conditional on p, her credence in $p \wedge q$ is 0.7, so she prefers to believe it. So conditionalising on p does change her preferences with respect to believing $p \wedge q$, so she doesn't believe p. So the effect of these stipulations rules out much more than just belief in propositions whose credence is below 1/2.

This suggests the third, and most important point. The problem with the threshold view was that it led to violations of closure. Given the theory as stated, we can prove the following theorem. Whenever p and q and their conjunction are all open or salient, and both are believed, and the agent is probabilistically coherent, the agent also believes $p \wedge q$. This is a quite restricted closure principle, but this is no reason to deny that it is *true*, as it fails to be true on the threshold view.

The proof of this theorem is a little complicated, but worth working through. First we'll prove that if the agent believes p, believes q, and p and q are both salient, then the agent prefers believing $p \wedge q$ to not believing it, if $p \wedge q$ is eligible for belief. In what follows $Pr(x|y)$ is the agent's conditional degree of belief in x given y. Since the agent is coherent, we'll assume this is a probability function (hence the name).

1. Since the agent believes that q, they prefer believing that q to not believing that q (by the criteria for belief)
2. So the agent prefers believing that q to not believing that q given p (From 1 and the fact that they believe that p, and that q is salient)
3. So $Pr(q|p) > 1/2$ (from 2)
4. $Pr(q|p) = Pr(p \wedge q|p)$ (by probability calculus)
5. So $Pr(p \wedge q|p) > 1/2$ (from 3, 4)
6. So, if $p \wedge q$ is eligible for belief, then the agent prefers believing that $p \wedge q$ to not believing it, given p (from 5)
7. So, if $p \wedge q$ is eligible for belief, the agent prefers believing that $p \wedge q$ to not believing it (from 6, and the fact that they believe that p, and $p \wedge q$ is salient)

So whenever, p, q and $p \wedge q$ are salient, and the agent believes each conjunct, the agent prefers believing the conjunction $p \wedge q$ to not believing it, if $p \wedge q$ is eligible. Now we have to prove that $p \wedge q$ is eligible for belief, to prove that it is actually believed. That is, we have to prove that (5) follows from (4) and (3), where the initial quantifiers range over actions that are open and salient *tout court*.

(3) $\forall \mathrm{A} \forall \mathrm{B} \forall r\ (\mathrm{A} \geq_r \mathrm{B} \leftrightarrow \mathrm{A} \geq_p \wedge r\ \mathrm{B})$
(4) $\forall \mathrm{A} \forall \mathrm{B} \forall r\ (\mathrm{A} \geq_r \mathrm{B} \leftrightarrow \mathrm{A} \geq_q \wedge r\ \mathrm{B})$
(5) $\forall \mathrm{A} \forall \mathrm{B} \forall r\ (\mathrm{A} \geq_r \mathrm{B} \leftrightarrow \mathrm{A} \geq_{p \wedge q \wedge r} \mathrm{B})$

Assume that (5) isn't true. That is, there are A, B and s such that $\neg(\mathrm{A} \geq_s \mathrm{B} \leftrightarrow \mathrm{A} \geq_{p \wedge q \wedge s} \mathrm{B})$. By hypothesis s is active, and consistent with $p \wedge q$. So it is the conjunction of relevant, salient propositions. Since q is salient, this means $q \wedge s$ is also active. Since s is consistent with $p \wedge q$, it follows that $q \wedge s$ is consistent with p. So $q \wedge s$ is a possible substitution instance for r in (3). Since (3) is true, it follows that $\mathrm{A} \geq_{q \wedge s} \mathrm{B} \leftrightarrow \mathrm{A} \geq_{p \wedge q \wedge s} \mathrm{B}$. By similar reasoning, it follows that s is a permissible substitution instance in (4), giving us $\mathrm{A} \geq_s \mathrm{B} \leftrightarrow \mathrm{A} \geq_{q \wedge s} \mathrm{B}$. Putting the last two biconditionals together we get $\mathrm{A} \geq_s \mathrm{B} \leftrightarrow \mathrm{A} \geq_{p \wedge q \wedge s} \mathrm{B}$, contradicting our hypothesis that there is a counterexample to (5). So whenever (3) and (4) are true, (5) is true as well, assuming p, q and $p \wedge q$ are all salient.

2.3 The Interest-Relativity of Belief

2.3.1 Interests and Functional Roles

The previous section was largely devoted to proving an existential claim: there is *some* interest-relativity to knowledge. Or, if you prefer, it proved a negative claim: the best theory of knowledge is *not* interest-neutral. But this negative conclusion invites a philosophical challenge: what is the best explanation of the interest-relativity of knowledge? My answer is in two parts. Part of the interest-relativity of knowledge comes from the interest-relativity of belief, and part of it comes from the fact that interests generate certain kinds of doxastic defeaters. It's the second part, the part that is new to this paper, that makes the theory a version of non-doxastic IRI.

Here's my theory of belief. *S* believes that p iff conditionalising on p doesn't change *S*'s answer to any relevant question. I'm using 'relevance' here in a non-technical sense; I say a lot more about how to cash out the notion in my (2005a). The key thing to note is that relevance is interest-relative, so the theory of belief is interest-relative. There is a bit more to say about what kind of *questions* are important for this definition of belief. In part because I've changed my mind a little bit on this since the earlier paper, I'll spend a bit more time on it. The following four kinds of questions are the most important.

- How probable is q?
- Is q or r more probable?
- How good an idea is it to do φ?
- Is it better to do φ or ψ?

The theory of belief says that someone who believes that $p$doesn't change their answer to any of these questions upon conditionalising on p. Putting this formally, and making the restriction to relevant questions explicit, we get the following theorems of our theory of belief.[4]

BAP For all relevant q, x, if p is believed then $\Pr(q) = x$ iff $\Pr(q|p) = x$.
BCP For all relevant q, r, if p is believed then $\Pr(q) \geq \Pr(r)$ iff $\Pr(q|p) \geq \Pr(r|p)$.
BAU For all relevant φ, x, if p is believed then $U(\varphi) = x$ iff $U(\varphi|p) = x$.
BCU For all relevant φ, ψ, if p is believed then $U(\varphi) \geq U(\psi)$ iff $U(\varphi|p) \geq U(\psi|p)$.

In the earlier paper I focussed on **BAU** and **BCU**. But **BAP** and **BCP** are important as well. Indeed, focussing on them lets us derive a nice result.

Charlie is trying to figure out exactly what the probability of p is. That is, for any $x \in [0, 1]$, whether $\Pr(p) = x$ is a relevant question. Now Charlie is well aware that $\Pr(p|p) = 1$. So unless $\Pr(p) = 1$, Charlie will give a different answer to the questions *How probable is p?* and *Given p, how probable is p?*. So unless Charlie holds that $\Pr(p)$ is 1, she won't count as believing that p. One consequence of this is that Charlie can't reason, "The probability of p is exactly 0.978, so p." That's all to the good, since that looks like bad reasoning. And it looks like bad reasoning even though in some circumstances Charlie can rationally believe propositions that she (rationally) gives credence 0.978 to. Indeed, in some circumstances she can rationally believe something *in virtue* of it being 0.978 probable.

That's because the reasoning in the previous paragraph assumes that every question of the form *Is the probability of p equal to x?* is relevant. In practice, fewer

[4] In the last two lines, I use $U(\varphi)$ to denote the expected utility of φ, and $U(\varphi|p)$ to denote the expected utility of φ conditional on p. It's often easier to write this as simply $U(\varphi \wedge p)$, since the utility of φ conditional on p just is the utility of doing φ in a world where p is true. That is, it is the utility of $\varphi \wedge p$ being realised. But we get a nicer symmetry between the probabilistic principles and the utility principles if we use the explictly conditional notation for each.

questions than that will be relevant. Let's say that the only questions relevant to Charlie are of the form *What is the probability of p to one decimal place?*. And assume that no other questions become relevant in the course of her inquiry into this question.[5] Charlie decides that to the first decimal place, $\Pr(p) = 1.0$, i.e., $\Pr(p) > 0.95$. That is compatible with simply believing that p. And that seems right; if for practical purposes, the probability of p is indistinguishable from 1, then the agent is confident enough in p to believe it.

So there are some nice features of this theory of belief. Indeed, there are several reasons to believe it. It is, I have argued, the best functionalist account of belief. I'm not going to argue for functionalism about the mind, since the argument would take at least a book. (The book in question might look a lot like Braddon-Mitchell and Jackson (2007).) But I do think functionalism is true, and so the best functionalist theory of belief is the best theory of belief.

The argument for this theory of belief in my (2005a) rested heavily on the flaws of rival theories. We can see those flaws by looking at a tension that any theory of the relationship between belief and credence must overcome. Each of the following three principles seems to be plausible.

1. If S has a greater credence in p than in q, and she believes q, then she believes p as well; and if her credences in both p and q are rational, and her belief in q is rational, then so is her belief in p.
2. If S rationally believes p and rationally believes q, then it is open to her to rationally believe $p \wedge q$ without changing her credences.
3. S can rationally believe p while having credence of less than 1 in p.

But these three principles, together with some principles that are genuinely uncontroversial, entail an absurd result. By 3, there is some p such that $Cr(p) = x < 1$, and p is believed. (Cr is the function from any proposition to our agent's credence in that propositions.) Let S know that a particular fair lottery has l tickets, where $l > {}^{1}/_{1-x}$. The uncontroversial principle we'll use is that in such a case S's credence that any given ticket will lose should be ${}^{l-1}/_{l}$. Since ${}^{l-1}/_{l} > x$, it follows by 1 that S believes of each ticket that it will lose. Since her credences are rational, these beliefs are rational. By repeated applications of 2 then, the agent can rationally believe that each ticket will lose. But she rationally gives credence

[5] This is probably somewhat unrealistic. It's hard to think about whether $\Pr(p)$ is closer to 0.7 or 0.8 without raising to salience questions about, for example, what the second decimal place in $\Pr(p)$ is. This is worth bearing in mind when coming up with intuitions about the cases in this paragraph.

0 to the proposition that each ticket will lose. So by 1 she can rationally believe any proposition in which her credence is greater than 0. This is absurd.[6]

I won't repeat all the gory details here, but one of the consequences of the discussion in Weatherson (2005a) was that we could hold on to 3, and to restricted versions of 1 and 2. In particular, if we restricted 1 and 2 to relevant propositions (in some sense) they became true, although the unrestricted version is false. A key part of the argument of the earlier paper was that this was a better option than the more commonly taken option of holding on to unrestricted versions of 1 and 3, at the cost of abandoning 2 even in clear cases. But one might wonder why I'm holding so tightly on to 3. After all, there is a functionalist argument that 3 is false.

A key functional role of credences is that if an agent has credence x in p she should be prepared to buy a bet that returns 1 util if p, and 0 utils otherwise, iff the price is no greater than p utils. A key functional role of belief is that if an agent believes p, and recognises that φ is the best thing to do given p, then she'll do φ. Given p, it's worth paying any price up to 1 util for a bet that pays 1 util if p. So believing p seems to mean being in a functional state that is like having credence 1 in p.

But this argument isn't quite right. If we spell out more carefully what the functional roles of credence and belief are, a loophole emerges in the argument that belief implies credence 1. The interest-relative theory of belief turns out to exploit that loophole. What's the difference, in functional terms, between having credence x in p, and having credence $x + \varepsilon$ in p? Well, think again about the bet that pays 1 util if p, and 0 utils otherwise. And imagine that bet is offered for $x + \varepsilon/2$ utils. The person whose credence is x will decline the offer; the person whose credence is $x + \varepsilon$ will accept it. Now it will usually be that no such bet is on offer.[7] No matter; as long as one agent is *disposed* to accept the offer, and the other agent is not, that suffices for a difference in credence.

The upshot of that is that differences in credences might be, indeed usually will be, constituted by differences in dispositions concerning how to act in choice situations far removed from actuality. I'm not usually in a position of having to accept or decline a chance to buy a bet for 0.9932 utils that the local coffee shop is currently open. Yet whether I would accept or decline such a bet matters

[6] See Sturgeon (2008) for discussion of a similar puzzle for anyone trying to tell a unified story of belief and credence.

[7] There are exceptions, especially in cases where p concerns something significant to financial markets, and the agent trades financial products. If you work through the theory that I'm about to lay out, one consequence is that such agents should have very few unconditional beliefs about financially-sensitive information, just higher and lower credences. I think that's actually quite a nice outcome, but I'm not going to rely on that in the argument for the view.

to whether my credence that the coffee shop is open is 0.9931 or 0.9933. This isn't a problem with the standard picture of how credences work. It's just an observation that the high level of detail embedded in the picture relies on taking the constituents of mental states to involve many dispositions.

One of the crucial features of the theory of belief I'm defending is that what an agent believes is in general *insensitive* to such abtruse dispositions, although it is very sensitive to dispositions about practical matters. It's true that if I believe that p, and I'm rational enough, I'll act as if p is true. Is it also true that if I believe p, I'm disposed to act as if p is true no matter what choices are placed in front of me? The theory being defended here says no, and that seems plausible. As we say in the case of Barry and Beth, Barry can believe that p, but be disposed to *lose that belief* rather than act on it if odd choices, like that presented by the genie, emerge.

This suggests the key difference between belief and credence 1. For a rational agent, a credence of 1 in p means that the agent is disposed to answer a wide range of questions the same way she would answer that question conditional on p. That follows from the fact that these four principles are trivial theorems of the orthodox theory of expected utility.[8]

C1AP For all q, x, if $\Pr(p) = 1$ then $\Pr(q) = x$ iff $\Pr(q|p) = x$.
C1CP For all q, r, if $\Pr(p) = 1$ then $\Pr(q) \geq \Pr(r)$ iff $\Pr(q|p) \geq \Pr(r|p)$.
C1AU For all φ, x, if $\Pr(p) = 1$ then $U(\varphi) = x$ iff $U(\varphi|p) = x$.
C1CP For all φ, ψ, if $\Pr(p) = 1$ then $U(\varphi) \geq U(\psi)$ iff $U(\varphi|p) \geq U(\psi|p)$.

Those look a lot like the theorems of the theory of belief that we discussed above. But note that these claims are *unrestricted*, whereas in the theory of belief, we restricted attention to relevant actions, propositions, utilities and probabilities. That turns out to be the difference between belief and credence 1. Since that difference is interest-relative, belief is interest-relative.

I used to think that that was all the interest-relativity we needed in epistemology. Now I don't, for reasons that I'll go through in section three. (Readers who care more about the theory of knowledge than the theory of belief may want to skip ahead to that section.) But first I want to clean up some loose ends in the acount of belief.

[8] The presentation in this section, as in the earlier paper, assumes at least a weak form of consequentialism in the sense of Hammond (1988). This was arguably a weakness of the earlier paper. We'll return to the issue of what happens in cases where the agent doesn't, and perhaps shouldn't, maximise expected utility, at the end of the section.

2.3.2 Two Caveats

The theory sketched so far seems to me right in the vast majority of cases. It fits in well with a broadly functionalist view of the mind, and it handles difficult cases, like that of Charlie, nicely. But it needs to be supplemented and clarified a little to handle some other difficult cases. In this section I'm going to supplement the theory a little to handle what I call 'impractical propositions', and say a little about morally loaded action.

Jones has a false geographic belief: he believes that Los Angeles is west of Reno, Nevada.[9] This isn't because he's ever thought about the question. Rather, he's just disposed to say "Of course" if someone asks, "Is Los Angeles west of Reno?" That disposition has never been triggered, because no one's ever bothered to ask him this. Call the proposition that Los Angeles is west of Reno p.

The theory given so far will get the right result here: Jones does believe that p. But it gets the right answer for an odd reason. Jones, it turns out, has very little interest in American geography right now. He's a schoolboy in St Andrews, Scotland, getting ready for school and worried about missing his schoolbus. There's no inquiry he's currently engaged in for which p is even close to relevant. So conditionalising on p doesn't change the answer to any inquiry he's engaged in, but that would be true no matter what his credence in p is.

There's an immediate problem here. Jones believes p, since conditionalising on p doesn't change the answer to any relevant inquiry. But for the very same reason, conditionalising on $\neg p$ doesn't change the answer to any relevant inquiry. It seems our theory has the bizarre result that Jones believes $\neg p$ as well. That is both wrong and unfair. We end up attributing inconsistent beliefs to Jones simply because he's a harried schoolboy who isn't currently concerned with the finer points of geography of the American southwest.

Here's a way out of this problem in four relatively easy steps.[10] First, we say that which questions are relevant questions is not just relative to the agent's interests, but also relevant to the proposition being considered. A question may be relevant relative to p, but not relative to q. Second, we say that relative to p, the question of whether p is more probable than $\neg p$ is a relevant question. Third, we infer from that that an agent only believes p if their credence in p is greater than their credence in $\neg p$, i.e., if their credence in p is greater than 1/2. Finally, we say that when the issue is whether the subject believes that p, the question of whether p is more probable than $\neg p$ is not only relevant on its own, but it stays

[9] I'm borrowing this example from Fred Dretske, who uses it to make some interesting points about dispositional belief.

[10] The recipe here is similar to that given in Weatherson (2005a), but the motivation is streamlined. Thanks to Jacob Ross for helpful suggestions here.

being a relevant question conditional on any q that is relevant to the subject. In the earlier paper (Weatherson, 2005a) I argue that this solves the problem raised by impractical propositions in a smooth and principled way.

That's the first caveat. The second is one that isn't discussed in the earlier paper. If the agent is merely trying to get the best outcome for themselves, then it makes sense to represent them as a utility maximiser. And within orthodox decision theory, it is easy enough to talk about, and reason about, conditional utilities. That's important, because conditional utilities play an important role in the theory of belief offered at the start of this section. But if the agent faces moral constraints on her decision, it isn't always so easy to think about conditional utilities.

When agents have to make decisions that might involve them causing harm to others if certain propositions turn out to be true, then I think it is best to supplement orthodox decision theory with an extra assumption. The assumption is, roughly, that for choices that may harm others, expected value is absolute value. It's easiest to see what this means using a simple case of three-way choice. The kind of example I'm considering here has been used for (slightly) different purposes by Frank Jackson (1991).

The agent has to do φ or ψ. Failure to do either of these will lead to disaster, and is clearly unacceptable. Either φ or ψ will avert the disaster, but one of them will be moderately harmful and the other one will not. The agent has time before the disaster to find out which of φ and ψ is harmful and which is not for a nominal cost. Right now, her credence that φ is the harmful one is, quite reasonably, 1/2. So the agent has three choices:

- Do φ;
- Do ψ; or
- Wait and find out which one is not harmful, and do it.

We'll assume that other choices, like letting the disaster happen, or finding out which one is harmful and doing it, are simply out of consideration. In any case, they are clearly dominated options, so the agent shouldn't do them. Let p be the propostion that φ is the harmful one. Then if we assume the harm in question has a disutility of 10, and the disutility of waiting to act until we know which is the harmful one is 1, the values of the possible outcomes are as follows:

	p	$\neg p$
Do φ	-10	0
Do ψ	0	-10
Find out which is harmful	-1	-1

Given that $Pr(p) = 1/2$, it's easy to compute that the expected value of doing either φ or ψ is -5, while the expected value of finding out which is harmful is -1, so the agent should find out which thing is to be done before acting. So far most consequentialists would agree, and so probably would most non-consequentialists for most ways of fleshing out the abstract example I've described.[11]

But most consequentialists would also say something else about the example that I think is not exactly true. Just focus on the column in the table above where p is true. In that column, the highest value, 0, is alongside the action *Do* ψ. So you might think that conditional on p, the agent should do ψ. That is, you might think the conditional expected value of doing ψ, conditional on p being true, is 0, and that's higher than the conditional expected value of any other act, conditional on p. If you thought that, you'd certainly be in agreement with the orthodox decision-theoretic treatment of this problem.

In the abstract statement of the situation above, I said that one of the options would be *harmful*, but I didn't say who it would be harmful to. I think this matters. I think what I called the orthodox treatment of the situation is correct when the harm accrues to the person making the decision. But when the harm accrues to another person, particularly when it accrues to a person that the agent has a duty of care towards, then I think the orthodox treatment isn't quite right.

My reasons for this go back to Jackson's original discussion of the puzzle. Let the agent be a doctor, the actions φ and ψ be her prescribing different medications to a patient, and the harm a severe allergic reaction that the patient will have to one of the medications. Assume that she can run a test that will tell her which medication the patient is allergic to, but the test will take a day. Assume that the patient will die in a month without either medication; that's the disaster that must be averted. And assume that the patient is is some discomfort that either medication would relieve; that's the small cost of finding out which medication is the risk. Assume finally that there is no chance the patient will die in the day it takes to run the test, so the cost of running the test really is nominal.

A good doctor in that situation will find out which medication the patient is allergic to before prescribing either medicine. It would be *reckless* to prescribe a medicine that is unnecessary and that the patient might be allergic to. It is worse than reckless if the patient is actually allergic to the medicine prescribed, and the doctor harms the patient. But even if she's lucky and prescribes the 'right' medication, the recklessness remains. It was still, it seems, the wrong thing for her to do.

[11] Some consequentialists say that what the agent should do depends on whether p is true. If p is true, she should do ψ, and if p is false she should do φ. As we'll see, I have reasons for thinking this is rather radically wrong.

All of that is in Jackson's discussion of the case, though I'm not sure he'd agree with the way I'm about to incorporate these ideas into the formal decision theory. Even under the assumption that p, prescribing ψ is still wrong, because it is reckless. That should be incorporated into the values we ascribe to different actions in different circumstances. The way I do it is to associate the value of each action, in each circumstance, with its actual expected value. So the decision table for the doctor's decision looks something like this.

	p	$\neg p$
Do φ	-5	-5
Do ψ	-5	-5
Find out which is harmful	-1	-1

In fact, the doctor is making a decision under certainty. She knows that the value of prescribing either medicine is -5, and the value of running the tests is -1, so she should run the tests.

In general, when an agent has a duty to maximise the expected value of some quantity Q, then the value that goes into the agent's decision table in a cell is *not* the value of Q in the world-action pair the agent represents. Rather, it's the expected value of Q given that world-action pair. In situations like this one where the relevant facts (e.g., which medicine the patient is allergic to) don't affect the evidence the agent has, the decision is a decision under *certainty*. This is all as things should be. When you have obligations that are drawn in terms of the expected value of a variable, the actual values of that variable cease to be directly relevant to the decision problem.

Similar morals carry across to theories that offer a smaller role to expected utility in determining moral value. In particular, it's often true that decisions where it is uncertain what course of action will produce the best outcome might still, in the morally salient sense, be decisions under certainty. That's because the uncertainty doesn't impact how we should weight the different possible outcomes, as in orthodox utility theory, but how we should value them. That's roughly what I think is going on in cases like this one, which Jessica Brown has argued are problematic for the epistemological theories John Hawthorne and Jason Stanley have recently been defending.[12]

[12] The target here is not directly the interest-relativity of their theories, but more general principles about the role of knowledge in action and assertion. Since my theories are close enough, at least in consequences, to Hawthorne and Stanley's, it is important to note how my theory handles the case.

> A student is spending the day shadowing a surgeon. In the morning he observes her in clinic examining patient A who has a diseased left kidney. The decision is taken to remove it that afternoon. Later, the student observes the surgeon in theatre where patient A is lying anaesthetised on the operating table. The operation hasn't started as the surgeon is consulting the patient's notes. The student is puzzled and asks one of the nurses what's going on:
>
> **Student**: I don't understand. Why is she looking at the patient's records? She was in clinic with the patient this morning. Doesn't she even know which kidney it is?
>
> **Nurse**: Of course, she knows which kidney it is. But, imagine what it would be like if she removed the wrong kidney. She shouldn't operate before checking the patient's records. (Brown, 2008, 1144-1145)

It is tempting, but for reasons I've been going through here mistaken, to represent the surgeon's choice as follows. Let **Left** mean the left kidney is diseased, and **Right** mean the right kidney is diseased.

	Left	**Right**
Remove left kidney	1	-1
Remove right kidney	-1	1
Check notes	$1-\varepsilon$	$1-\varepsilon$

Here ε is the trivial but non-zero cost of checking the chart. Given this table, we might reason that since the surgeon knows that she's in the left column, and removing the left kidney is the best option in that column, she should remove the left kidney rather than checking the notes.

But that reasoning assumes that the surgeon does not have any epistemic obligations over and above her duty to maximise expected utility. And that's very implausible. It's totally implausible on a non-consequentialist moral theory. A non-consequentialist may think that some people have just the same obligations that the consequentialist says they have – legislators are frequently mentioned as an example – but surely they wouldn't think *surgeons* are in this category. And even a consequentialist who thinks that surgeons have special obligations in terms of their institutional role should think that the surgeon's obligations go above and beyond the obligation every agent has to maximise expected utility.

It's not clear exactly what the obligation the surgeon has. Perhaps it is an obligation to not just know which kidney to remove, but to know this on the basis of evidence she has obtained while in the operating theatre. Or perhaps it is an obligation to make her belief about which kidney to remove as sensitive as possible to various possible scenarios. Before she checked the chart, this counterfactual was false: *Had she misremembered which kidney was to be removed, she would have a true belief about which kidney was to be removed.* Checking the chart makes that counterfactual true, and so makes her belief that the left kidney is to be removed a little more sensitive to counterfactual possibilities.

However we spell out the obligation, it is plausible given what the nurse says that the surgeon has some such obligation. And it is plausible that the 'cost' of violating this obligation, call it δ is greater than the cost of checking the notes. So here is the decision table the surgeon faces.

	Left	**Right**
Remove left kidney	$1-\delta$	$-1-\delta$
Remove right kidney	$-1-\delta$	$1-\delta$
Check notes	$1-\varepsilon$	$1-\varepsilon$

And it isn't surprising, or a problem for an interest-relative theory of knowledge or belief, that the surgeon should check the notes, even if she believes *and knows* that the left kidney is the diseased one.

2.4 *Playing Games with a Lockean*

I'm going to raise problems for Lockeans, and for defenders of regularity in general, by discussing a simple game. 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 it is different to what I used to think, and perhaps to what some other people think too, so I'll set it out slowly.[13]

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 able.

[13] 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.

	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 that we discussed a couple of paragraphs ago. The relevant states of the world are choices of Column. But that's fine; we didn't say in setting out the decision problem what constituted the states S_1 and S_2. And note that we solved the problem without explicitly saying anything about probabilities. What we added was some information about Column's payouts, and the fact that Column is rational. From there we deduced something about Column's play, namely that she would play S_2. And from that we concluded what Row should 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. (We'll come back to whether the 'mostly' qualification is needed below.) 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, we can't say much about a decision problem with unspecified probabilities. But we can and do say a lot about games where the setup of the game doesn't specify the probabilities, because we can solve for them given the other information we have.

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. Arguably the core solution concept in game theory is the Nash equilibrium. As you probably know, 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 puzzles. 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.[14] 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 RED-GREEN. I've said something different about this game in earlier work (Weatherson, 2012a). But I now think that to understand what's going on, we need to think about mixed strategies where one element of the mixture has probability one.

Informally, in this game A and B must each play either a green or red card. I will capitalise A's moves, i.e., A can play GREEN or RED, and italicise B's moves, i.e., B can play *green* or *red*. If two green cards, or one green card and one red card are played, each player gets \$1. If two red cards 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.

[14] Actually, I guess 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.

	green	*red*
GREEN	1, 1	1, 1
RED	1, 1	0, 0

When I write game tables like this, and I think this is the usual way game tables are to be interpreted (Weatherson, 2012b), 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. 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. We'll relax this later, but it will help the discussion a lot I think 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 (White, 2005). I don't, but it does seem like it might *often* hold. 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 GREEN should be the same as the probability of B playing *green*, 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 GREEN, is 1. Indeed, the known value of GREEN is 1. If the probability of *green* is x, then the expected value of RED is x. So if $x < 1$, and Row is rational, she'll definitely play GREEN. But that's inconsistent with the claim that $x < 1$, since that means that it isn't definite that Row will play GREEN.

So we can conclude that $x = 1$. Does that mean we can know that Row will play GREEN? 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 *green*. Since any rational person can conclude this, Row can conclude it. So Row knows that she'll get 1 whether she plays GREEN or RED. But then she should be indifferent between playing GREEN and RED. And if we know she's indifferent between playing GREEN and RED, 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 GREEN.

I think the arguments of the last two paragraphs are sound. We'll turn to an objection presently, but let's note how bizarre is the conclusion we've reached. One argument has shown that it could not be more probable that Row will play GREEN. A second argument has shown that we can't know that Row will play GREEN. It reminds me of examples involving blindspots (Sorensen, 1988). Consider this case:

(B) Brian does not know (B).

That's true, right? Assume it's false, so I do know (B). Knowledge is factive, so (B) is true. But that contradicts the assumption that it's false. So it's true. But I obviously don't know that it's true; that's what this very true proposition says.

Now I'm not going to rest anything on this case, because there are so many tricky things one can say about blindspots, and about the paradoxes generally. It does suggest that there are other finite cases where one can properly have maximal credence in a true proposition without knowledge.[15] And, assuming that we shouldn't believe things we know we don't know, that means we can have maximal credence in things we don't believe. All I want to point out is that this phenomena of maximal credence without knowledge, and presumably without full belief, isn't a quirky feature of self-reference, or of games, or of puzzles about infinity; it comes up in a wide range of cases.

For the rest of this section 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 GREEN, we could show that *A* had no reason to play GREEN. But actually what we showed was that the expected utility of playing GREEN would be the same as playing RED. Perhaps *A* has a reason to play GREEN, namely that GREEN weakly dominates RED. After all, there's one possibility on the table where GREEN does better than RED, 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 GREEN 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

[15] As an aside, the existence of these cases is why I get so irritated when epistemologists try to theorise about 'Gettier Cases' as a class. What does (B) have in common with inferences from a justified false belief, or with otherwise sound reasoning that is ever so close to issuing in a false conclusion due to relatively bad luck? As far as I can tell, the class of justified true beliefs that aren't knowledge is a disjunctive mess, and this should matter for thinking about the nature of knowledge. For further examples, see Williamson (2012).

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 GREEN/*green*, 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 *blue*, and if *B* plays *blue*, *A* will receive 2 if she plays RED, and -1 if she plays GREEN. Well hold on, you might think, didn't I say that *green* and *red* 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, GREEN does better than RED.

We can put this as a dilemma. Assume, for *reductio*, that GREEN/*green* is the only rational play. Then if we restrict our attention to possibilities that are epistemically open to *A*, then GREEN does just as well as RED; 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 RED. Since *A* maximises expected utility, that means *A* must have credence 1 that *B* will play *green*. 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 *green*, it isn't obvious that playing RED is rational. After all, relative to the space of epistemic possibilities, GREEN weakly dominates RED. 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 RED. 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 GREEN, like whether the coin will ever land tails, is a case the Lockean cannot get right, no matter where they set the threshold for belief; our credence is above the threshold, but we don't believe.

So I think this case is a real problem for a Lockean view about the relationship between credence and belief. If *A* is rational, she can have credence 1 that *B* will play *green*, but won't believe that *B* will play *green*. But now you might worry that my own account of the relationship between belief and credence is in just as much trouble. After all, I said that to believe p is, roughly, to have the same attitudes towards all salient questions as you have conditional on p. And it's hard to identify a question that rational *A* would answer differently upon conditionalising on the proposition that *B* plays *green*.

I think what went wrong in my earlier view was that I'd too quickly equated updating with conditionalisation. The two can come apart. Here's an example from Gillies (2010) that makes the point well.

> 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)

As Gillies goes on to point out, this isn't really a problem for the Ramsey test view of conditionals.

> 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.

For example, in the game I've been discussing, we should believe that rational *B* might play *red*. Indeed, the probability of that is, I think, 1. And whether or not *B* might play red is highly salient; it matters to the probability of whether **A** will play GREEN or RED. Conditionalising on something that has probability 1, such as that *B* will play *green*, can hardly change that probability. But updating on the proposition that *B* will play *green* can make a difference. We can see that by simply noting that the conditional *If B plays green, she might play red* is incoherent.

So I conclude that a theory of belief like mine can handle the puzzle this game poses, as long as it distinguishes between conditionalising and updating, in just the way Gillies suggests. To believe that *p* is to be disposed to not change any attitude towards a salient question on updating that *p*. (Plus some bells and whistles to deal with propositions that are not relevant to salient questions. We'll return to them below.) Updating often goes by conditionalisation, so we can often say that belief means having attitudes that match unconditionally and conditionally on *p*. But not all updating works that way, and the theory of belief needs to acknowledge this.

2.5 The Power of Theoretical Interests

So I think we should accept that credences exist. And we can just about reduce beliefs to credences. In previous work I argued that we could do such a reduction. I'm not altogether sure whether the amendments to that view I'm proposing here means it no longer should count as a reductive view; we'll come back to that question in the conclusion.

The view I defended in previous work is that the reduction comes through the relationship between conditional and unconditional attitudes. Very roughly, to believe that *p* is simply to have the same attitudes, towards all salient questions, unconditionally as you have conditional on *p*. In a syrupy slogan, belief means never having to say you've conditionalised. For reasons I mentioned in section 1, I now think that was inaccurate; I should have said that belief means never having to say you've updated, or at least that you've updated your view on any salient question.

The restriction to salient questions is important. Consider any p that I normally take for granted, but such that I wouldn't bet on it at insane odds. I prefer declining such a bet to taking it. But conditional on p, I prefer taking the bet. So that means I don't believe any such p. But just about any p satisfies that description, for at least some 'insane' odds. So I believe almost nothing. That would be a *reductio* of the position. I respond by saying that the choice of whether to take an insane bet is not normally salient.

But now there's a worry that I've let in too much. For many p, there is no salient decision that they even bear on. What I would do conditional on p, conditional on $\neg p$, and unconditionally is exactly the same, over the space of salient choices. (And this isn't a case where updating and conditionalising come apart; I'll leave this proviso mostly implicit from now on.) So with the restriction in place, I believe p and $\neg p$. That seems like a *reductio* of the view too. I probably do have inconsistent beliefs, but not in virtue of p being irrelevant to me right now. I've changed my mind a little about what the right way to avoid this problem is, in part because of some arguments by Jacob Ross and Mark Schroeder.

They have what looks like, on the surface, a rather different view to mine. They say that to believe p is to have a **default reasoning disposition** to use p in reasoning. Here's how they describe their own 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.

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

I think they take all three of these points to be reasons to favour their view over mine. As I see it, we agree on point 1 (and I always had the resources to agree with them), I can accommodate point 2 with a modification to my theory, and point 3 is a cost of their theory, not a benefit. Let's take those points in order.

There are lots of reasons to dislike what Ross and Schroeder call *Pragmatic Credal Reductionism* (PCR). This is, more or less, the view that the salient questions, in the sense relevant above, are just those which are practically relevant to the agent. So to believe p just is to have the same attitude towards all practically relevant questions unconditionally as conditional on p. There are at least three reasons to resist this view.

One reason comes from the discussions of Ned Block's example Blockhead (Block, 1978). As Braddon-Mitchell and Jackson point out, the lesson to take from that example is that beliefs are constituted in part by their relations to other mental states (Braddon-Mitchell and Jackson, 2007, 114ff). There's a quick attempted refutation of PCR via the Blockhead case which doesn't quite work. We might worry that if all that matters to belief given PCR is how it relates to action, PCR will have the implausible consequence that Blockhead has a rich set of beliefs. That isn't right; PCR is compatible with the view that Blockhead doesn't have credences, and hence doesn't have credences that constitute beliefs. But the Blockhead examples value isn't exhausted by its use in quick refutations.[16] The

[16] The point I'm making here is relevant I think to recent debates about the proper way to formalise counterexamples in philosophy, as in (Williamson, 2007; Ichikawa and Jarvis, 2009; Malmgren, 2011). I worry that too much of that debate is focussed on the role that examples play in one-step refutations. But there's more, much more, to a good example than that.

lesson is that beliefs are, by their nature, interactive. It seems to me that PCR doesn't really appreciate that lesson.

Another reason comes from recent work by Jessica Brown (forthcoming). Compare these two situations.

1. *S* is in circumstances *C*, and has to decide whether to do *X*.
2. *S* is in completely different circumstances to *C*, but is seriously engaged in planning for future contingencies. She's currently trying to decide whether in circumstances *C* to do *X*.

Intuitively, *S* can bring exactly the same evidence, knowledge and beliefs to bear on the two problems. If *C* is a particularly high stakes situation, say it is a situation where one has to decide what to feed someone with a severe peanut allergy, then a lot of things that can ordinarily be taken for granted cannot, in this case, be taken for granted. And that's true whether *S* is actually in *C*, or she is just planning for the possibility that she finds herself in *C*.

So I conclude that both practical and theoretical interests matter for what we can take for granted in inquiry. The things we can take for granted into a theoretical inquiry into what to do in high stakes contexts as restricted, just as they are when we are in a high stakes context, and must make a practical decision. Since the latter restriction on what we can take for granted is explained by (and possibly constituted by) a restriction on what we actually believe in those contexts, we should similarly conclude that agents simply believe less when they are reasoning about high stakes contexts, whatever their actual context.

A third reason to dislike PCR comes from the 'Renzi' example in Ross and Schroeder's paper. I'll run through a somewhat more abstract version of the case, because I don't think the details are 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 with all this.

They also think that PCR implies that the agent *does* believe q. The reason is that conditionalising on q doesn't change the agent's beliefs about any practical question. I think that's right too, at least on a natural understanding of what 'practical' is.

My response to all these worries is to say that whether someone believes that p depends not just on how conditionalising (or more generally updating) on p would affect someone's action, but on how it would affect their reasoning. That is, just as we learned from the Blockhead example, to believe that p requires having a mental state that is connected to the rest of one's cognitive life in roughly the way a belief that p should be connected. Let's go through both the last two cases to see how this works on my theory.

One of the things that happens when the stakes go up is that conditionalising on very probable things can change the outcome of interesting decisions. Make the probability that some nice food is peanut-free be high, but short of one. Conditional on it being peanut-free, it's a good thing to give to a peanut-allergic guest. But unconditionally, it's a bad thing to give to such a guest, because the niceness of the food doesn't outweigh the risk of killing them. And that's true whether the guest is actually there, or you're just thinking about what to do should such a guest arrive in the future. In general, the same questions will be relevant whether you're in C trying to decide whether to do X, or simply trying to decide whether to X in C. In one case they will be practically relevant questions, in the other they will be theoretically relevant questions. But this feels a lot like a distinction without a difference, since the agent should have similar beliefs in the two cases.

The same response works for Ross and Schroeder's case. 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. (In (Weatherson, 2005a) I said that to justifiably believe q was to have a justified credence in q that was sufficiently high to count as a belief. The considerations of the last two sentences puts some pressure on that reductive theory of justification for beliefs.)

So I think that Ross and Schroeder and I agree on point 1; something beyond practical interests is relevant to belief.

They have another case that I think does suggest a needed revision to my theory. I'm going to modify their case a little to change the focus a little, and to avoid puzzles about vagueness. (What follows is a version of their example about Dalí's moustache, purged of any worries about vagueness, and without the focus on consistency. I don't think the problem they true to press on me, that my theory allows excessive inconsistency of belief among rational agents, really sticks. Everyone will have to make qualifications to consistency to deal with the preface paradox, and for reasons I went over in (Weatherson, 2005a), I think the qualifications I make are the best ones to make.)

Let D be the proposition that the number of games the Detroit Tigers won in 1976 (in the MLB regular season) is not a multiple of 3. At most times, D is completely irrelevant to anything I care about, either practically or theoretically. My attitudes towards any relevant question are the same unconditionally as conditional on D. So there's a worry that I'll count as believing D, and believing $\neg D$, by default.

In earlier work, I added a clause meant to help with cases like this. I said that for determining whether an agent believes that p, we should treat the question of whether p's probability is above or below 0.5 as salient, even if the agent doesn't care about it. Obviously this won't help with this particular case. The probability of D is around 2/3, and is certainly above 0.5. My 'fix' avoids the consequence that I implausibly count as believing $\neg D$. But I still count, almost as implausibly, as believing D. This needs to be fixed.

Here's my proposed change. For an agent to count as believing p, it must be possible for p to do some work for them in reasoning. Here's what I mean by work. Consider a very abstract set up of a decision problem, as follows.

	p	q
X	4	1
Y	3	2

That table encodes a lot of information. It encodes that $p \vee q$ is true; otherwise there are some columns missing. It encodes that the only live choices are X or Y; otherwise there are rows missing. It encodes that doing X is better than doing Y if p, and worse if q.

For any agent, and any decision problem, there is a table like this that they would be disposed to use to resolve that problem. Or, perhaps, there are a series of tables and there is no fact about which of them they would be most disposed to use.

Given all that terminology, here's my extra constraint on belief. To believe that p, there must be some decision problem such that some table the agent would be disposed to use to solve it encodes that p. If there is no such problem, the agent does not believe that p. For anything that I intuitively believe, this is an easy condition to satisfy. Let the problem be whether to take a bet that pays 1 if p, and loses 1 otherwise. Here's the table I'd be disposed to use to solve the problem.

	p
Take bet	1
Decline bet	0

This table encodes that p, so it is sufficient to count as believing that p. And it doesn't matter that this bet isn't on the table. I'm disposed to use this table, so that's all that matters.

But might there be problems in the other direction. What about an agent who, if offered such a bet on D, would use such a simple table? I simply say that they believe that D. I would not use any such table. I'd use this table.

	D	$\neg D$
Take bet	1	−1
Decline bet	0	0

Now given the probability of D, I'd still end up taking the bet; it has an expected return of $2/3$. (Well, actually I'd probably decline the bet because being offered the bet would change the probability of D for reasons made clear in (Runyon, 1992, 14–15). But that hardly undermines the point I'm making.) But this isn't some analytic fact about me, or even I think some respect in which I'm obeying the dictates of rationality. It's simply a fact that I wouldn't take D for granted in any inquiry. And that's what my non-belief that D consists in.

This way of responding to the Tigers example helps respond to a nice observation that Ross and Schroeder make about correctness. A belief that p is, in some sense, *incorrect* if $\neg p$. It isn't altogether clear how to capture this sense given a simple reduction of beliefs to credences. I propose to capture it using this idea that decision tables encode propositions. A table is incorrect if it encodes something that's false. To believe something is, *inter alia*, to be disposed to use a table that encodes it. So if it is false, it involves a disposition to do something incorrect.

It also helps capture Holton's observation that beliefs should be resilient. If someone is disposed to use decision tables that encode that *p*, that disposition should be fairly resilient. And to the extent that it is resilient, they will satisfy all the other clauses in my preferred account of belief. So anyone who believes *p* should have a resilient belief that *p*.

The last point is where I think my biggest disagreement with Ross and Schroeder lies. 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. (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 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. (22)

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, reopen 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 we've stated 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.

CHAPTER 3

CONJUNCTIONS AND PREFACES

3.1 Defending Closure

So on my account of the connection between degrees of belief and belief *tout court*, probabilistic coherence implies logical coherence amongst salient propositions. The last qualification is necessary. It is possible for a probabilistically coherent agent to not believe the *non*-salient consequences of things they believe, and even for a probabilistically coherent agent to have inconsistent beliefs as long as not all the members of the inconsistent set are active. Some people argue that even this weak a closure principle is implausible. David Christensen (2005), for example, argues that the preface paradox provides a reason for doubting that beliefs must be closed under entailment, or even must be consistent. Here is his description of the case.

> 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. Although he does not say this, nothing in his discussion suggests that he is using the irrelevance of some of the propositions in the author's defence. So here is an argument that we should abandon closure amongst relevant beliefs.

Christensen's discussion, like other discussions of the preface paradox, makes frequent use of the fact that examples like these are quite common. We don't have to go to fake barn country to find a counterexample to closure. 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 in forthcoming work by Ishani Maitra (Maitra, 2010), and is briefly mentioned by Christensen in setting out the problem. We only have a counterexample to closure if the author *believes* every thing she writes in her book. (Indeed, we only have a counterexample if she reasonably believes every one of them. But we'll assume a rational author who only believes what she ought to believe.) This seems unlikely to be true to me. 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. It seems to me (following here Maitra's discussion) that 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.

There are three objections I know to this picture, none of them particularly conclusive. First, Christensen says that the author doesn't qualify their assertions. But neither does our detective qualify most individual sentences. Second, Christensen says 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 reporting her hypothesis. Third, Timothy Williamson (2000) has argued that if speakers don't believe what they say, we won't have an explanation of why Moore's paradoxical sentences, like "The butler did it, but I don't believe the butler did it," are always defective. Whatever the explanation of the paradoxicality of these sentences might be, the alleged requirement that speakers believe what they say can't be it. For our detective cannot properly say "The butler did it, but I don't believe the butler did it" in setting out her hypothesis, even though *believing* the butler did it is not necessary for her to say "The butler did it" in setting out just that hypothesis.

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 in his paper "Believing Conjunctions". The following situation does not involve me believing anything inconsistent.

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

If we read the first claim *de dicto*, that I believe that Manny just said something false, then there is no inconsistency. (Unless I also believe 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 I believe to be false, then the situation does involve me 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 section started, let alone this paper.) 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)

I'm not sure this is undoubtedly true; it isn't clear that propositions (as opposed to their representations) have lengths. And humans can believe propositions that *can* be represented by sentences as long as books. But even without that point, Christensen is right that there is an idealisation here, since ordinary humans do not know exactly what is in a given book, and hence don't have *de re* attitudes towards the propositions expressed in the book.

I'm actually rather suspicious of the intuition that Christensen is pushing here, that idealising in this way doesn't change intuitions about the case. 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, I find the intuitions become rather murky. But I won't press this point.

A more important point is that we believers in closure don't think that authors should think their books are inerrant. Rather, following Stalnaker (1984), we think that authors shouldn't unqualifiedly *believe* the individual statements in their book if they don't believe the conjunction of those statements. Rather, their attitude towards those propositions (or at least some of them) should be that they are probably true. (As Stalnaker puts it, they accept the story without believing it.) Proponents of the preface paradox know that this is a possible response, and tend to argue 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 closure. 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.

a) They are part of a research community where only asserting propositions you believe is compatible with active scholarship;
b) 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
c) 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 seems just absurd.

I'm not entirely sure an agent subject to exactly these constraints is even fully conceivable. (Such an agent is negatively conceivable, in David Chalmers's terminology, but I rather doubt they are positively conceivable.) But even if they are a genuine possibility, why the norms applicable to an agent satisfying that very gerrymandered set of constraints should be considered relevant norms for our state is far from clear. I'd go so far as to say it's clear that the applicability (or otherwise) of a given norm to such an odd agent is no reason whatsoever to say it applies to us. But since the preface paradox only provides a reason for just these kinds of agents to violate closure, we have no reason for ordinary humans to violate closure. So I see no reason here to say that we can have probabilistic coherence without logical coherence, as proponents of the threshold view insist we can have, but which I say we can't have *at least when the propositions involved are salient*. The more pressing question, given the failure of the preface paradox argument, is why I don't endorse a much stronger closure principle, one that drops the restriction to salient propositions. The next section will discuss that point.

I've used Christensen's book as a stalking horse in this section, because it is the clearest and best statement of the preface paradox. Since Christensen is a paradox-mongerer and I'm a paradox-denier, it might be thought we have a deep disagreement about the relevant epistemological issues. But actually I think our overall views are fairly close despite this. I favour an epistemological outlook I call "Probability First", the view that getting the epistemology of partial belief right is of the first importance, and everything else should flow from that. Christensen's view, reduced to a slogan, is "Probability First and Last". This section has been basically about the difference between those two slogans. It's an important dispute, but it's worth bearing in mind that it's a factional squabble within the Probability Party, not an outbreak of partisan warfare.

3.2 Too Little Closure?

In the previous section I defended the view that a coherent agent has beliefs that are deductively cogent with respect to salient propositions. Here I want to defend the importance of the qualification. Let's start with what I take to be the most important argument for closure, the passage from Stalnaker's *Inquiry* that I quoted above.

> 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 threshold view, as long as the threshold 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 threshold view, though it seems absurd at first that this could be coherent.

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 salient. And in that case closure, deductive coherence, 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.

The position I endorse here is very similar in its conclusions to that endorsed by Gilbert Harman 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.)

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

This seems right to me, both what it says and its implicature that the reason in question is not a conclusive reason. My main objection to those who use the preface paradox to argue against closure is that they give us a mistaken picture of what we have *to do* epistemically. When I have inconsistent beliefs, or I don't believe some consequence of my beliefs, that is something I have a reason to deal with at some stage, something I have to do. 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 bathroom, yet here I am writing this paper, and (given the relevant deadlines) 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 bathroom) isn't something to be done. The bathroom 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. My attic is to be cleaned, or 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 I haven't followed through that it couldn't possibly be worth my time to sort 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 garbage. 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 metaphor that Gilbert Ryle develops in *The Concept of Mind* at a point where he's covering 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.

For a less metaphorical reason for doubting the wisdom of this unchecked commitment to closure, we might notice the difficulties theorists tend to get into all sorts of difficulties. Consider, for example, the view put forward by Mark Kaplan in *Decision Theory as Philosophy*. Here is his definition of belief.

> You count as believing P just if, were your sole aim to assert the truth (as it pertains to P), and you only options were to assert that P, assert that ¬P or make neither assertion, you would prefer to assert that P. (109)

Kaplan notes that conditional definitions like this are prone to Shope's conditional fallacy. If my sole aim were to assert the truth, I might have different beliefs to what I now have. He addresses one version of this objection (namely that it appears to imply that everyone believes their sole desire is to assert the truth) but as we'll see presently he can't avoid all versions of it.

These arguments are making me thirsty. I'd like a beer. Or at least I think I would. But wait! On Kaplan's theory I can't think that I'd like a beer, for if my sole aim were to assert the truth as it pertains to my beer-desires, I wouldn't have beer desires. And then I'd prefer to assert that I wouldn't like a beer, I'd merely like to assert the truth as it pertains to my beer desires.

Even bracketing this concern, Kaplan ends up being committed to the view that I can (coherently!) believe that *p* even while regarding *p* as highly improbable. This looks like a refutation of the view to me, but Kaplan accepts it with some equanimity. He has two primary reasons for saying we should live with this. First, he says that it only looks like an absurd consequence if we are committed to the Threshold View. To this all I can say is that *I* don't believe the Threshold View, but it still seems absurd to me. Second, he says that any view is going to have to be revisionary to some extent, because our ordinary concept of belief is not "coherent" (142). His view is that, "Our ordinary notion of belief both construes belief as a state of confidence short of certainty and takes consistency of belief to be something that is at least possible and, perhaps, even desirable" and this is impossible. I think the view here interprets belief as a state less than confidence and allows for as much consistency as the folk view does (i.e. consistency amongst salient propositions), so this defence is unsuccessful as well.

None of the arguments here in favour of our restrictions on closure are completely conclusive. In part the argument at this stage rests on the lack of a plausible rival theory that doesn't interpret belief as certainty but implements a stronger closure principle. It's possible that tomorrow someone will come up with a theory that does just this. Until then, we'll stick with the account here, and see what its epistemological implications might be.

3.3 Examples of Pragmatic Encroachment

Fantl and McGrath's case for pragmatic encroachment starts with cases like the following. (The following case is not quite theirs, but is similar enough to suit their plan, and easier to explain in my framework.)

> *Local and Express*
>
> There are two kinds of trains that run from the city to the suburbs: the local, which stops at all stations, and the express, which skips the first eight stations. Harry and Louise want to go to the fifth station, so they shouldn't catch the Express. Though if they do it isn't too hard to catch a local back the other way, so it isn't usually a large cost. Unfortunately, the trains are not always clearly labelled. They see a particular train about to leave. If it's a local they are better off catching it, if it is an express they should wait for the next local, which they can see is already boarding passengers and will leave in

> a few minutes. While running towards the train, they hear a fellow passenger say "It's a local." This gives them good, but far from overwhelming, reason to believe that the train is a local. Passengers get this kind of thing wrong fairly frequently, but they don't have time to get more information. So each of them face a gamble, which they can take by getting on the train. If the train is a local, they will get home a few minutes early. If it is an express they will get home a few minutes later. For Louise, this is a low stakes gamble, as nothing much turns on whether she is a few minutes early or late, but she does have a weak preference for arriving earlier rather than later. But for Harry it is a high stakes gamble, because if he is late he won't make the start of his daughter's soccer game, which will highly upset her. There is no large payoff for Harry arriving early.

What should each of them do? What should each of them believe?

The first question is relatively easy. Louise should catch the train, and Harry should wait for the next. For each of them that's the utility maximising thing to do. The second one is harder. Fantl and McGrath suggest that, despite being in the same epistemic position with respect to everything except their interests, Louise is justified in believing the train is a local and Harry is not. I agree. (If you don't think the particular case fits this pattern, feel free to modify it so the difference in interests grounds a difference in what they are justified in believing.) Does this show that our notion of epistemic justification has to be pragmatically sensitive? I'll argue that it does not.

The fundamental assumption I'm making is that what is primarily subject to epistemic evaluation are degrees of belief, or what are more commonly called states of confidence in ordinary language. When we think about things this way, we see that Louise and Harry are justified in adopting *the very same degrees of belief*. Both of them should be confident, but not absolutely certain, that the train is a local. We don't have even the appearance of a counterxample to Probabilistic Evidentialism here. If we like putting this in numerical terms, we could say that each of them is justified in assigning a probability of around 0.9 to the proposition *That train is a local*.[1] So as long as we adopt a Probability First epistemology, where we in the first instance evaluate the probabilities that agents assign to propositions, Harry and Louise are evaluated alike iff they do the same thing.

[1] I think putting things numerically is misleading because it suggests that the kind of bets we usually use to measure degrees of belief are open, salient options for Louise and Harry. But if those bets were open and salient, they wouldn't *believe* the train is a local. Using qualitative rather than quantitative language to describe them is just as accurate, and doesn't have misleading implications about their practical environment.

How then can we say that Louise alone is justified in believing that the train is a local? Because that state of confidence they are justified in adopting, the state of being fairly confident but not absolutely certain that the train is a local, counts as believing that the train is a local given Louise's context but not Harry's context. Once Louise hears the other passenger's comment, conditionalising on *That's a local* doesn't change any of her preferences over open, salient actions, including such 'actions' as believing or disbelieving propositions. But conditional on the train being a local, Harry prefers catching the train, which he actually does not prefer.

In cases like this, interests matter not because they affect the degree of confidence that an agent can reasonably have in a proposition's truth. (That is, not because they matter to epistemology.) Rather, interests matter because they affect whether those reasonable degrees of confidence amount to belief. (That is, because they matter to philosophy of mind.) There is no reason here to let pragmatic concerns into epistemology.

Chapter 4

Beliefs and Action

4.1 *Examples of Pragmatic Encroachment*

Fantl and McGrath's case for pragmatic encroachment starts with cases like the following. (The following case is not quite theirs, but is similar enough to suit their plan, and easier to explain in my framework.)

> *Local and Express*
>
> There are two kinds of trains that run from the city to the suburbs: the local, which stops at all stations, and the express, which skips the first eight stations. Harry and Louise want to go to the fifth station, so they shouldn't catch the Express. Though if they do it isn't too hard to catch a local back the other way, so it isn't usually a large cost. Unfortunately, the trains are not always clearly labelled. They see a particular train about to leave. If it's a local they are better off catching it, if it is an express they should wait for the next local, which they can see is already boarding passengers and will leave in a few minutes. While running towards the train, they hear a fellow passenger say "It's a local." This gives them good, but far from overwhelming, reason to believe that the train is a local. Passengers get this kind of thing wrong fairly frequently, but they don't have time to get more information. So each of them face a gamble, which they can take by getting on the train. If the train is a local, they will get home a few minutes early. If it is an express they will get home a few minutes later. For Louise, this is a low stakes gamble, as nothing much turns on whether she is a few minutes early or late, but she does have a weak preference for arriving earlier rather than later. But for Harry it is a high stakes gamble, because if he is late he won't make the start of his daughter's soccer game, which will highly upset her. There is no large payoff for Harry arriving early.

What should each of them do? What should each of them believe?

The first question is relatively easy. Louise should catch the train, and Harry should wait for the next. For each of them that's the utility maximising thing to do. The second one is harder. Fantl and McGrath suggest that, despite being in the same epistemic position with respect to everything except their interests, Louise is justified in believing the train is a local and Harry is not. I agree. (If you don't think the particular case fits this pattern, feel free to modify it so the difference in interests grounds a difference in what they are justified in believing.) Does this show that our notion of epistemic justification has to be pragmatically sensitive? I'll argue that it does not.

The fundamental assumption I'm making is that what is primarily subject to epistemic evaluation are degrees of belief, or what are more commonly called states of confidence in ordinary language. When we think about things this way, we see that Louise and Harry are justified in adopting *the very same degrees of belief*. Both of them should be confident, but not absolutely certain, that the train is a local. We don't have even the appearance of a counterxample to Probabilistic Evidentialism here. If we like putting this in numerical terms, we could say that each of them is justified in assigning a probability of around 0.9 to the proposition *That train is a local*.[1] So as long as we adopt a Probability First epistemology, where we in the first instance evaluate the probabilities that agents assign to propositions, Harry and Louise are evaluated alike iff they do the same thing.

How then can we say that Louise alone is justified in believing that the train is a local? Because that state of confidence they are justified in adopting, the state of being fairly confident but not absolutely certain that the train is a local, counts as believing that the train is a local given Louise's context but not Harry's context. Once Louise hears the other passenger's comment, conditionalising on *That's a local* doesn't change any of her preferences over open, salient actions, including such 'actions' as believing or disbelieving propositions. But conditional on the train being a local, Harry prefers catching the train, which he actually does not prefer.

In cases like this, interests matter not because they affect the degree of confidence that an agent can reasonably have in a proposition's truth. (That is, not because they matter to epistemology.) Rather, interests matter because they affect whether those reasonable degrees of confidence amount to belief. (That is,

[1] I think putting things numerically is misleading because it suggests that the kind of bets we usually use to measure degrees of belief are open, salient options for Louise and Harry. But if those bets were open and salient, they wouldn't *believe* the train is a local. Using qualitative rather than quantitative language to describe them is just as accurate, and doesn't have misleading implications about their practical environment.

because they matter to philosophy of mind.) There is no reason here to let pragmatic concerns into epistemology.

4.2 Justification and Practical Reasoning

The discussion in the last section obviously didn't show that there is no encroachment of pragmatics into epistemology. There are, in particular, two kinds of concerns one might have about the prospects for extending my style of argument to block all attempts at pragmatic encroachment. The biggest concern is that it might turn out to be impossible to defend a Probability First epistemology, particularly if we do not allow ourselves pragmatic concerns. For instance, it is crucial to this project that we have a notion of evidence that is not defined in terms of traditional epistemic concepts (e.g. as knowledge), or in terms of interests. This is an enormous project, and I'm not going to attempt to tackle it here. The second concern is that we won't be able to generalise the discussion of that example to explain the plausibility of (JP) without conceding something to the defenders of pragmatic encroachment.

(JP) If *S* justifiably believes that *p*, then *S* is justified in using *p* as a premise in practical reasoning.

And that's what we will look at in this section. To start, we need to clarify exactly what (JP) means. Much of this discussion will be indebted to Fantl and McGrath's discussion of various ways of making (JP) more precise. To see some of the complications at issue, consider a simple case of a bet on a reasonably well established historical proposition. The agent has a lot of evidence that supports *p*, and is offered a bet that returns $1 if *p* is true, and loses $500 if *p* is false. Since her evidence doesn't support *that* much confidence in *p*, she properly declines the bet. One might try to reason intuitively as follows. Assume that she justifiably believed that *p*. Then she'd be in a position to make the following argument.

> *p*
> If *p*, then I should take the bet
> So, I should take the bet

Since she isn't in a position to draw the conclusion, she must not be in a position to endorse both of the premises. Hence (arguably) she isn't justified in believing that *p*. But we have to be careful here. If we assume also that *p* is true (as Fantl and McGrath do, because they are mostly concerned with knowledge rather than justified belief), then the second premise is clearly false, since it is a conditional with a true antecedent and a false consequent. So the fact that she can't draw the conclusion of this argument only shows that she can't endorse *both* of the premises, and that's not surprising since one of the premises is most likely false. (I'm not assuming here that the conditional is true iff it has a true antecendent or a false consequent, just that it is only true if it has a false antecedent or a true consequent.)

In order to get around this problem, Fantl and McGrath suggest a few other ways that our agent might reason to the bet. They suggest each of the following principles.

> S knows that p only if, for any act A, if *S* knows that if p, then A is the best thing she can do, then *S* is rational to do A. (72)
>
> *S* knows that p only if, for any states of affairs A and B, if *S* knows that if p, then A is better for her than B, then *S* is rational to prefer A to B. (74)
>
> **(PC)** *S* is justified in believing that p only if *S* is rational to prefer as if p. (77)

Hawthorne (2004, 174-181) appears to endorse the second of these principles. He considers an agent who endorses the following implication concerning a proposed sell of a lottery ticket for a cent, which is well below its actuarially fair value.

> I will lose the lottery.
> If I keep the ticket, I will get nothing.
> If I sell the ticket, I will get a cent.
> So I ought to sell the ticket. (174)

(To make this fully explicit, it helps to add the tacit premise that a cent is better than nothing.) Hawthorne says that this is intuitively a *bad* argument, and concludes that the agent who attempts to use it is not in a position to know its first premise. But that conclusion only follows if we assume that the argument form is acceptable. So it is plausible to conclude that he endorses Fantl and McGrath's second principle.

The interesting question here is whether the theory endorsed in this paper can validate the true principles that Fantl and McGrath articulate. (Or, more precisely, we can validate the equivalent true principles concerning justified belief, since knowledge is outside the scope of the paper.) I'll argue that it can in the following way. First, I'll just note that given the fact that the theory here implies the closure principles we outlined in section 5, we can easily enough endorse Fantl and McGrath's first two principles. This is good, since they seem true. The longer part of the argument involves arguing that their principle (PC), which doesn't hold on the theory endorsed here, is in fact incorrect.

One might worry that the qualification on the closure principles in section 5 mean that we can't fully endorse the principles Fantl and McGrath endorse. In particular, it might be worried that there could be an agent who believes that *p*, believes that if *p*, then A is better than B, but doesn't put these two beliefs together to infer that A is better than B. This is certainly a possibility given the qualifications listed above. But note that in this position, if those two beliefs were justified, the agent would certainly be *rational* to conclude that A is better than B, and hence rational to prefer A to B. So the constraints on the closure principles don't affect our ability to endorse these two principles.

The real issue is (PC). Fantl and McGrath offer a lot of cases where (PC) holds, as well as arguing that it is plausibly true given the role of implications in practical reasoning. What's at issue is that (PC) is stronger than a deductive closure principle. It is, in effect, equivalent to endorsing the following schema as a valid principle of implication.

p
Given *p*, A is preferable to B
So, A is preferable to B

I call this Practical Modus Ponens, or PMP. The middle premise in PMP is *not* a conditional. It is not to be read as *If p, then A is preferable to B*. Conditional valuations are not conditionals. To see this, again consider the proposed bet on (true) *p* at exorbitant odds, where A is the act of taking the bet, and B the act of declining the bet. It's true that given *p*, A is preferable to B. But it's not true that if *p*, then A is preferable to B. Even if we restrict our attention to cases where the preferences in question are perfectly valid, this is a case where PMP is invalid. Both premises are true, and the conclusion is false. It might nevertheless be true that whenever an agent is justified in believing both of the premises, she is justified in believing the conclusion. To argue against this, we need a *very* complicated case, involving embedded bets and three separate agents, Quentin, Robby and Thom. All of them have received the same evidence, and all of them are faced with the same complex bet, with the following properties.

- *p* is an historical proposition that is well (but not conclusively) supported by their evidence, and happens to be true. All the agents have a high credence in *p*, which is exactly what the evidence supports.
- The bet A, which they are offered, wins if *p* is true, and loses if *p* is false.
- If they win the bet, the prize is the bet B.

- s is also an historical proposition, but the evidence tells equally for and against it. All the agents regard s as being about as likely as not. Moreover, s turns out to be false.
- The bet B is worth \$2 if s is true, and worth -\$1 if s is false. Although it is actually a losing bet, the agents all rationally value it at around 50 cents.
- How much A costs is determined by which proposition from the partition $\{q, r, s\}$ is true.
- If q is true, A costs \$2
- If r is true, A costs \$500
- If t is true, A costs \$1
- The evidence the agents has strongly supports r, though t is in fact true
- Quentin believes q
- Robby believes r
- Thom believes t

All of the agents make the utility calculations that their beliefs support, so Quentin and Thom take the bet and lose a dollar, while Robby declines it. Although Robby has a lot of evidence in favour of p, he correctly decides that it would be unwise to bet on p at effective odds of 1000 to 1 against. I'll now argue that both Quentin and Thom are potential counterexamples to (PC). There are three possibilities for what we can say about those two.

First, we could say that they are justified in believing p, and rational to take the bet. The problem with this position is that if they had rational beliefs about the partition $\{q, r, t\}$ they would realise that taking the bet does not maximise expected utility. If we take rational decisions to be those that maximise expected utility given a rational response to the evidence, then the decisions are clearly not rational.

Second, we could say that although Quentin and Thom are not rational in accepting the bet, nor are they justified in believing that p. This doesn't seem particularly plausible for several reasons. The irrationality in their belief systems concerns whether q, r or t is true, not whether p is true. If Thom suddenly got a lot of evidence that t is true, then all of his (salient) beliefs would be well supported by the evidence. But it is bizarre to think that whether his belief in p is rational turns on how much evidence he has for t. Finally, even if we accept that agents in higher stakes situations need more evidence to have justified beliefs, the fact is that the agents are in a low-risk situation, since t is actually true, so the most they could lose is \$1.

So it seems like the natural thing to say is that Quentin and Thom *are* justified in believing that p, and are justified in believing that given p, it maximises expected utility to take the bet, but they are not rational to take the bet. (At least, in the version of the story where they are thinking about which of q, r and t are correct given their evidence when thinking about whether to take the bet they are counterexamples to (PC).) Against this, one might respond that if belief in p is justified, there are arguments one might make to the conclusion that the bet should be taken. So it is inconsistent to say that the belief is justified, but the decision to take the bet is not rational. The problem is finding a premise that goes along with p to get the conclusion that taking the bet is rational. Let's look at some of the premises the agent might use.

- If p, then the best thing to do is to take the bet.

This isn't true (p is true, but the best thing to do isn't to take the bet). More importantly, the agents think this is only true if s is true, and they think s is a 50/50 proposition. So they don't believe this premise, and it would not be rational to believe it.

- If p, then probably the best thing to do is to take the bet.

Again this isn't true, and it isn't well supported, and it doesn't even support the conclusion, for it doesn't follow from the fact that x is probably the best thing to do that x should be done.

- If p, then taking the bet maximises rational expected utility.

This isn't true – it is a conditional with a true antecedent and a false consequent. Moreover, if Quentin and Thom were rational, like Robby, they would recognise this.

- If p, then taking the bet maximises expected utility relative to their beliefs.

This is true, and even reasonable to believe, but it doesn't imply that they should take the bet. It doesn't follow from the fact that doing something maximises expected utility relative to my crazy beliefs that I should do that thing.

- Given p, taking the bet maximises rational expected utility.

This is true, and even reasonable to believe, but it isn't clear that it supports the conclusion that the agents should take the bet. The implication appealed to here is PMP, and in this context that's close enough to equivalent to (PC). If we think that this case is a prima facie problem for (PC), as I think is intuitively plausible, then we can't use (PC) to show that it *doesn't* post a problem. We could obviously continue for a while, but it should be clear it will be very hard to find a way to justify taking the bet even spotting the agents *p* as a premise they can use in rational deliberation. So it seems to me that (PC) is not in general true, which is good because as we'll see in cases like this one the theory outlined here does not support it.

The theory we have been working with says that belief that *p* is justified iff the agent's degree of belief in *p* is sufficient to amount to belief in their context, and they are justified in believing *p* to that degree. Since by hypothesis Quentin and Thom are justified in believing *p* to the degree that they do, the only question left is whether this amounts to belief. This turns out not to be settled by the details of the case as yet specified. At first glance, assuming there are no other relevant decisions, we might think they believe that *p* because (a) they prefer (in the relevant sense) believing *p* to not believing *p*, and (b) conditionalising on *p* doesn't change their attitude towards the bet. (They prefer taking the bet to declining it, both unconditionally and conditional on *p*.)

But that isn't all there is to the definition of belief *tout court*. We must also ask whether conditionalising on *p* changes any preferences conditional on any active proposition. And that may well be true. Conditional on *r*, Quentin and Thom prefer not taking the bet to taking it. But conditional on *r* and *p*, they prefer taking the bet to not taking it. So if *r* is an active proposition, they don't believe that *p*. If *r* is not active, they do believe it. In more colloquial terms, if they are concerned about the possible truth of *r* (if it is salient, or at least not taken for granted to be false) then *p* becomes a potentially high-stakes proposition, so they don't believe it without extraordinary evidence (which they don't have). Hence they are only a counterexample to (PC) if *r* is not active. But if *r* is not active, our theory predicts that they are a counterexample to (PC), which is what we argued above is intuitively correct.

Still, the importance of *r* suggests a way of saving (PC). Above I relied on the position that if Quentin and Thom are not maximising rational expected utility, then they are being irrational. This is perhaps too harsh. There is a position we could take, derived from some suggestions made by Gilbert Harman in *Change in View*, that an agent can rationally rely on their beliefs, even if those beliefs were not rationally formed, if they cannot be expected to have kept track of the evidence they used to form that belief. If we adopt this view, then we might be able to say that (PC) is compatible with the correct normative judgments about this case.

To make this compatibility explicit, let's adjust the case so Quentin takes q for granted, and cannot be reasonably expected to have remembered the evidence for q. Thom, on the other hand, forms the belief that t rather than r is true in the course of thinking through his evidence that bears on the rationality of taking or declining the bet. (In more familiar terms, t is part of the inference Thom uses in coming to conclude that he should take the bet, though it is not part of the final implication he endorses whose conclusion is that he should take the bet.) Neither Quentin nor Thom is a counterexample to (PC) thus understood. (That is, with the notion of rationality in (PC) understood as Harman suggests that it should be.) Quentin is not a counterexample, because he is *rational* in taking the bet. And Thom is not a counterexample, because in his context, where r is active, his credence in p does not amount to belief in p, so he is not justified in believing p.

We have now two readings of (PC). On the strict reading, where a rational choice is one that maximises rational expected utility, the principle is subject to counterexample, and seems generally to be implausible. On the loose reading, where we allow agents to rely on beliefs formed irrationally in the past in rational decision making, (PC) *is* plausible. Happily, the theory sketched here agrees with (PC) on the plausible loose reading, but not on the implausible strict reading. In the previous section I argued that the theory also accounts for intuitions about particular cases like *Local and Express*. And now we've seen that the theory accounts for our considered opinions about which principles connecting justified belief to rational decision making we should endorse. So it seems at this stage that we can account for the intuitions behind the pragmatic encroachment view while keeping a concept of probabilistic epistemic justification that is free of pragmatic considerations.

4.3 Odds and Stakes

It is common to describe IRI as a theory where in 'high stakes' situations, more evidence is needed for knowledge than in 'low stakes' situations. But this is at best misleading. What really matters are the odds on any bet-like decision the agent faces with respect to the target proposition. More precisely, interests affect belief because whether someone believes p depends *inter alia* on whether their credence in p is high enough that any bet on p they actually face is a good bet. Raising the stakes of any bet on p does not directly change that, but changing the odds of the bets on p they face does change it. Now in practice due to the declining marginal utility of material goods, high stakes situations will usually be situations where an agent faces long odds. But it is the odds that matter to knowledge, not the stakes.

Some confusion on this point may have been caused by the Bank Cases that Stanley uses, and the Train Cases that Fantl and McGrath use, to motivate IRI. In those cases, the authors lengthen the odds the relevant agents face by increasing the potential losses the agent faces by getting the bet wrong. But we can make the same point by decreasing the amount the agent stands to gain by taking the bet. Let's go through a pair of cases, which I'll call the Map Cases, that illustrate this.

High Cost Map: Zeno is walking to the Mysterious Bookshop in lower Manhattan. He's pretty confident that it's on the corner of Warren Street and West Broadway. But he's been confused about this in the past, forgetting whether the east-west street is Warren or Murray, and whether the north-south street is Greenwich, West Broadway or Church. In fact he's right about the location this time, but he isn't justified in having a credence in his being correct greater than about 0.95. While he's walking there, he has two options. He could walk to where he thinks the shop is, and if it's not there walk around for a few minutes to the nearby corners to find where it is. Or he could call up directory assistance, pay $1, and be told where the shop is. Since he's confident he knows where the shop is, and there's little cost to spending a few minutes walking around if he's wrong, he doesn't do this, and walks directly to the shop.

Low Cost Map: Just like the previous case, except that Zeno has a new phone with more options. In particular, his new phone has a searchable map, so with a few clicks on the phone he can find where the store is. Using the phone has some very small costs. For example, it distracts him a little, which marginally raises the likelihood of bumping into another pedestrian. But the cost is very small compared to the cost of getting the location wrong. So even though he is very confident about where the shop is, he double checks while walking there.

I think the Map Cases are like the Bank Cases, Train Cases etc., in all important respects. I think Zeno knows where the shop is in High Cost Map, and doesn't know in Low Cost Map. And he doesn't know in Low Cost Map because the location of the shop has suddenly become the subject matter of a bet at very long odds. You should think of Zeno's not checking the location of the shop on his phone-map as a bet on the location of the shop. If he wins the bet, he wins a few seconds of undistracted strolling. If he loses, he has to walk around a few blocks looking for a store. The disutility of the loss seems easily twenty times greater than the utility of the gain, and by hypothesis the probability of winning the bet is no greater than 0.95. So he shouldn't take the bet. Yet if he knew where the store was, he would be justified in taking the bet. So he doesn't know where the store is. Now this is not a case where higher *stakes* defeat knowledge. If anything, the stakes are lower in Low Cost Map. But the relevant odds are longer, and that's what matters to knowledge.[2]

4.4 The Power of Theoretical Interests

So I think we should accept that credences exist. And we can just about reduce beliefs to credences. In previous work I argued that we could do such a reduction. I'm not altogether sure whether the amendments to that view I'm proposing here means it no longer should count as a reductive view; we'll come back to that question in the conclusion.

The view I defended in previous work is that the reduction comes through the relationship between conditional and unconditional attitudes. Very roughly, to believe that p is simply to have the same attitudes, towards all salient questions, unconditionally as you have conditional on p. In a syrupy slogan, belief means never having to say you've conditionalised. For reasons I mentioned in section 1, I now think that was inaccurate; I should have said that belief means never having to say you've updated, or at least that you've updated your view on any salient question.

The restriction to salient questions is important. Consider any p that I normally take for granted, but such that I wouldn't bet on it at insane odds. I prefer declining such a bet to taking it. But conditional on p, I prefer taking the bet. So that means I don't believe any such p. But just about any p satisfies that description, for at least some 'insane' odds. So I believe almost nothing. That would be a *reductio* of the position. I respond by saying that the choice of whether to take an insane bet is not normally salient.

But now there's a worry that I've let in too much. For many p, there is no salient decision that they even bear on. What I would do conditional on p, conditional on $\neg p$, and unconditionally is exactly the same, over the space of salient choices. (And this isn't a case where updating and conditionalising come apart; I'll leave this proviso mostly implicit from now on.) So with the restriction in place, I believe p and $\neg p$. That seems like a *reductio* of the view too.

[2] Note that I'm not claiming that it is intuitive that Zeno has knowledge in High Cost Map, but not that Low Cost Map. Nor am I claiming that we should believe IRI because it gets the Map Cases right. In fact, I don't believe either of those things. Instead I believe Zeno has knowledge in High Cost Map and not in Low Cost Map because I believe IRI is correct, and that's what IRI says about the case. It is sometimes assumed, e.g, in the experimental papers I'll discuss in section 7.3, that pairs of cases like these are meant to *motivate*, and not just *illustrate*, IRI. I can't speak for everyone's motivations, but I'm only using these cases as illustrations, not motivations.

I probably do have inconsistent beliefs, but not in virtue of *p* being irrelevant to me right now. I've changed my mind a little about what the right way to avoid this problem is, in part because of some arguments by Jacob Ross and Mark Schroeder.

They have what looks like, on the surface, a rather different view to mine. They say that to believe *p* is to have a **default reasoning disposition** to use *p* in reasoning. Here's how they describe their own 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.

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

I think they take all three of these points to be reasons to favour their view over mine. As I see it, we agree on point 1 (and I always had the resources to agree with them), I can accommodate point 2 with a modification to my theory, and point 3 is a cost of their theory, not a benefit. Let's take those points in order.

There are lots of reasons to dislike what Ross and Schroeder call *Pragmatic Credal Reductionism* (PCR). This is, more or less, the view that the salient questions, in the sense relevant above, are just those which are practically relevant to the agent. So to believe p just is to have the same attitude towards all practically relevant questions unconditionally as conditional on p. There are at least three reasons to resist this view.

One reason comes from the discussions of Ned Block's example Blockhead (Block, 1978). As Braddon-Mitchell and Jackson point out, the lesson to take from that example is that beliefs are constituted in part by their relations to other mental states (Braddon-Mitchell and Jackson, 2007, 114ff). There's a quick attempted refutation of PCR via the Blockhead case which doesn't quite work. We might worry that if all that matters to belief given PCR is how it relates to action, PCR will have the implausible consequence that Blockhead has a rich set of beliefs. That isn't right; PCR is compatible with the view that Blockhead doesn't have credences, and hence doesn't have credences that constitute beliefs. But the Blockhead examples value isn't exhausted by its use in quick refutations.[3] The lesson is that beliefs are, by their nature, interactive. It seems to me that PCR doesn't really appreciate that lesson.

Another reason comes from recent work by Jessica Brown (forthcoming). Compare these two situations.

1. *S* is in circumstances *C*, and has to decide whether to do *X*.
2. *S* is in completely different circumstances to *C*, but is seriously engaged in planning for future contingencies. She's currently trying to decide whether in circumstances *C* to do *X*.

Intuitively, *S* can bring exactly the same evidence, knowledge and beliefs to bear on the two problems. If *C* is a particularly high stakes situation, say it is a situation where one has to decide what to feed someone with a severe peanut allergy, then a lot of things that can ordinarily be taken for granted cannot, in this case, be taken for granted. And that's true whether *S* is actually in *C*, or she is just planning for the possibility that she finds herself in *C*.

[3] The point I'm making here is relevant I think to recent debates about the proper way to formalise counterexamples in philosophy, as in (Williamson, 2007; Ichikawa and Jarvis, 2009; Malmgren, 2011). I worry that too much of that debate is focussed on the role that examples play in one-step refutations. But there's more, much more, to a good example than that.

So I conclude that both practical and theoretical interests matter for what we can take for granted in inquiry. The things we can take for granted into a theoretical inquiry into what to do in high stakes contexts as restricted, just as they are when we are in a high stakes context, and must make a practical decision. Since the latter restriction on what we can take for granted is explained by (and possibly constituted by) a restriction on what we actually believe in those contexts, we should similarly conclude that agents simply believe less when they are reasoning about high stakes contexts, whatever their actual context.

A third reason to dislike PCR comes from the 'Renzi' example in Ross and Schroeder's paper. I'll run through a somewhat more abstract version of the case, because I don't think the details are 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 with all this.

They also think that PCR implies that the agent *does* believe q. The reason is that conditionalising on q doesn't change the agent's beliefs about any practical question. I think that's right too, at least on a natural understanding of what 'practical' is.

My response to all these worries is to say that whether someone believes that p depends not just on how conditionalising (or more generally updating) on p would affect someone's action, but on how it would affect their reasoning. That is, just as we learned from the Blockhead example, to believe that p requires having a mental state that is connected to the rest of one's cognitive life in roughly the way a belief that p should be connected. Let's go through both the last two cases to see how this works on my theory.

One of the things that happens when the stakes go up is that conditionalising on very probable things can change the outcome of interesting decisions. Make the probability that some nice food is peanut-free be high, but short of one. Conditional on it being peanut-free, it's a good thing to give to a peanut-allergic guest. But unconditionally, it's a bad thing to give to such a guest, because the niceness of the food doesn't outweigh the risk of killing them. And that's true whether the guest is actually there, or you're just thinking about what to do should such a guest arrive in the future. In general, the same questions will be relevant whether you're in C trying to decide whether to do X, or simply trying to decide whether to X in C. In one case they will be practically relevant questions, in the other they will be theoretically relevant questions. But this feels a lot like a distinction without a difference, since the agent should have similar beliefs in the two cases.

The same response works for Ross and Schroeder's case. 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. (In (Weatherson, 2005a) I said that to justifiably believe q was to have a justified credence in q that was sufficiently high to count as a belief. The considerations of the last two sentences puts some pressure on that reductive theory of justification for beliefs.)

So I think that Ross and Schroeder and I agree on point 1; something beyond practical interests is relevant to belief.

They have another case that I think does suggest a needed revision to my theory. I'm going to modify their case a little to change the focus a little, and to avoid puzzles about vagueness. (What follows is a version of their example about Dalí's moustache, purged of any worries about vagueness, and without the focus on consistency. I don't think the problem they true to press on me, that my theory allows excessive inconsistency of belief among rational agents, really sticks. Everyone will have to make qualifications to consistency to deal with the preface paradox, and for reasons I went over in (Weatherson, 2005a), I think the qualifications I make are the best ones to make.)

Let D be the proposition that the number of games the Detroit Tigers won in 1976 (in the MLB regular season) is not a multiple of 3. At most times, D is completely irrelevant to anything I care about, either practically or theoretically.

My attitudes towards any relevant question are the same unconditionally as conditional on D. So there's a worry that I'll count as believing D, and believing $\neg D$, by default.

In earlier work, I added a clause meant to help with cases like this. I said that for determining whether an agent believes that p, we should treat the question of whether p's probability is above or below 0.5 as salient, even if the agent doesn't care about it. Obviously this won't help with this particular case. The probability of D is around $2/3$, and is certainly above 0.5. My 'fix' avoids the consequence that I implausibly count as believing $\neg D$. But I still count, almost as implausibly, as believing D. This needs to be fixed.

Here's my proposed change. For an agent to count as believing p, it must be possible for p to do some work for them in reasoning. Here's what I mean by work. Consider a very abstract set up of a decision problem, as follows.

	p	q
X	4	1
Y	3	2

That table encodes a lot of information. It encodes that $p \vee q$ is true; otherwise there are some columns missing. It encodes that the only live choices are X or Y; otherwise there are rows missing. It encodes that doing X is better than doing Y if p, and worse if q.

For any agent, and any decision problem, there is a table like this that they would be disposed to use to resolve that problem. Or, perhaps, there are a series of tables and there is no fact about which of them they would be most disposed to use.

Given all that terminology, here's my extra constraint on belief. To believe that p, there must be some decision problem such that some table the agent would be disposed to use to solve it encodes that p. If there is no such problem, the agent does not believe that p. For anything that I intuitively believe, this is an easy condition to satisfy. Let the problem be whether to take a bet that pays 1 if p, and loses 1 otherwise. Here's the table I'd be disposed to use to solve the problem.

	p
Take bet	1
Decline bet	0

This table encodes that p, so it is sufficient to count as believing that p. And it doesn't matter that this bet isn't on the table. I'm disposed to use this table, so that's all that matters.

But might there be problems in the other direction. What about an agent who, if offered such a bet on D, would use such a simple table? I simply say that they believe that D. I would not use any such table. I'd use this table.

	D	$\neg D$
Take bet	1	−1
Decline bet	0	0

Now given the probability of D, I'd still end up taking the bet; it has an expected return of 2/3. (Well, actually I'd probably decline the bet because being offered the bet would change the probability of D for reasons made clear in (Runyon, 1992, 14–15). But that hardly undermines the point I'm making.) But this isn't some analytic fact about me, or even I think some respect in which I'm obeying the dictates of rationality. It's simply a fact that I wouldn't take D for granted in any inquiry. And that's what my non-belief that D consists in.

This way of responding to the Tigers example helps respond to a nice observation that Ross and Schroeder make about correctness. A belief that p is, in some sense, *incorrect* if $\neg p$. It isn't altogether clear how to capture this sense given a simple reduction of beliefs to credences. I propose to capture it using this idea that decision tables encode propositions. A table is incorrect if it encodes something that's false. To believe something is, *inter alia*, to be disposed to use a table that encodes it. So if it is false, it involves a disposition to do something incorrect.

It also helps capture Holton's observation that beliefs should be resilient. If someone is disposed to use decision tables that encode that p, that disposition should be fairly resilient. And to the extent that it is resilient, they will satisfy all the other clauses in my preferred account of belief. So anyone who believes p should have a resilient belief that p.

The last point is where I think my biggest disagreement with Ross and Schroeder lies. 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. (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 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. (22)

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 we've stated 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.

CHAPTER 5

GAMES, DECISIONS AND KNOWLEDGE

5.1 The Interest-Relativity of Knowledge

5.1.1 The Struction of Decision Problems

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 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 a little harder than it first seems 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, 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. 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]

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?

Let's start with a simpler problem that helps with both questions. Alice 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 Alice's story fairly schematic; let's fill in some of the details. Alice is on holiday at a casino. It's a fair casino; the probabilities of the outcomes of each of the games is just what you'd expect. And Alice 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

[1] If we are convinced that the right decision is the one that maximises expected utility, there is a sense in which these questions collapse. For the expected utility theorist, we can solve Dom's question by making sure the states are logically exhaustive, and making the 'payouts' in each state be expected payouts. But the theory that the correct decision is the one that maximises expected utility, while plausibly true, is controversial. It shouldn't be assumed when we are investigating the semantics of decision tables.

her token back and another token of the same value. There are many other bets she could make, but Alice 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.[2] 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 Alice 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 Alice doesn't believe the table reflects what will happen if she places the bet. As it turns out, Alice 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 Alice to put the token on the table. In fact, that is irrational. And it would still be irrational if Alice believes, *irrationally*, that the wheel will land on 28.

A better suggestion is that the table is misleading because Alice 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 seems it would be rational for her to put it on the table. If we take it as tacit 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 what the miscommunication between Professor Dec and her students was. 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.

1. 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

[2] Assuming Alice's utility curve for money curves downwards, she should be looking for a slightly higher chance of winning than 1/2 to place the bet, but that level of detail isn't relevant to the story we're telling here.

knows it to be true.

2. 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 Alice will get $1500 if C_2 and S_2) and implied by where the table ends (such as that S_3 will not happen). Could that relationship be weaker than knowledge? It's true that it is hard to come up with clear counterexamples to the suggestion that the relationship is merely justified true belief. But I think it is somewhat implausible to hold that the standard presentation of an example merely presupposes that the agent has a justified true belief that the table is correct, and does not in addition know that the table is correct.

My reasons for thinking this are similar to one of the reasons Timothy Williamson (Williamson, 2000, Ch. 9) gives for doubting that one's evidence is all that one justifiably truly believes. To put the point in Lewisian terms, it seems that knowledge is a much more *natural* relation than justified true belief. And when ascribing contents, especially contents of tacitly held beliefs, we should strongly prefer to ascribe more rather than less natural contents.[3]

[3] I'm here retracting some things I said a few years ago in a paper on philosophical methodology (Weatherson, 2003). There I argued that identifying knowledge with justified true belief would give us a theory on which knowledge was more natural than a theory on which we didn't identify knowledge with any other epistemic property. I now think that is wrong for a couple of reasons. First, although it's true (as I say in the earlier paper) that knowledge can't be primitive or perfectly natural, this doesn't make it less natural than justification, which is also far from a fundamental feature of reality. Indeed, given how usual it is for languages to have a simple representation of knowledge, we have some evidence that it is very natural for a term from a special science. Second, I think in the earlier paper I didn't fully appreciate the point (there attributed to Peter Klein) that the Gettier cases show that the property of being a justified true belief is not particularly natural. In general, when F and G are somewhat natural properties, then so is the property of being $F \wedge G$. But there are exceptions, especially in cases where these are properties that a whole can have in virtue of a part having the property. In those cases, a whole that has an F part and a G part will be $F \wedge G$, but this won't reflect any distinctive property of the whole. And one of the things the Gettier cases show is that the properties of *being justified* and *being true*, as applied to belief, fit this pattern.

Note that even if you think that philosophers are generally too quick to move from instinctive reactions to the Gettier case to abandoning the justified true belief theory of knowledge, this point holds up. What is important here is that on sufficient reflection, the Gettier cases show that some justified true beliefs are not knowledge, and that the cases in question also show that being a justified true belief is not a particularly natural or unified property. So the point I've been making in the last this footnote is independent of the point I wanted to stress in "What Good are Coun-

So the ‘special doxastic relationship’ is not weaker than knowledge. Could it be stronger? 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, I conclude that 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.

terexamples?”, namely, that philosophers in some areas (especially epistemology) are insufficiently reformist in their attitude towards our intuitive reactions to cases.

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 of 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 it wouldn't change what the agent should do if $\Pr(p)$ were raised to 1. 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, because of cases like the following.

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 Alice, 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 accepable 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 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 Alice 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.

5.1.2 From Decision Theory to Interest-Relativity

This way of thinking about decision problems offers a new perspective on the issue of whether we should always be prepared to bet on what we know.[4] To focus intuitions, let's take a concrete case. Barry is sitting in his apartment one evening when he hears a musician performing in the park outside. The musician, call her Beth, is one of Barry's favourite musicians, so the music is familiar to Barry. Barry is excited that Beth is performing in his neighbourhood, and he decides to hurry out to see the show. As he prepares to leave, a genie appears an offers him a bet.[5] If he takes the bet, and the musician is Beth, then the genie will give Barry ten dollars. On the other hand, if the musician is not Beth, he will be tortured in the fires of hell for a millenium. Let's put Barry's options in table form.

[4] This issue is of course central to the plotline in Hawthorne (2004).

[5] Assume, perhaps implausibly, that the sudden appearance of the genie is evidentially irrelevant to the proposition that the musician is Beth. The reasons this may be implausible are related to the arguments in (Runyon, 1992, 14-15). Thanks here to Jeremy Fantl.

	Musician is Beth	**Musician is not Beth**
Take Bet	Win $10	1000 years of torture
Decline Bet	Status quo	Status quo

Intuitively, it is extremely irrational for Barry to take the bet. People do make mistakes about identifying musicians, even very familiar musicians, by the strains of music that drift up from a park. It's not worth risking a millenium of torture for $10.

But it also seems that we've misstated the table. Before the genie showed up, it seemed clear that Barry knew that the musician was Beth. That was why he went out to see her perform. (If you don't think this is true, make the sounds from the park clearer, or make it that Barry had some prior evidence that Beth was performing which the sounds from the park remind him of. It shouldn't be too hard to come up with an evidential base such that (a) in normal circumstances we'd say Barry knew who was performing, but (b) he shouldn't take this genie's bet.) Now our decision tables should reflect the knowledge of the agent making the decision. If Barry knows that the musician is Beth, then the second column is one he knows will not obtain. So let's write the table in the standard form.

	Musician is Beth
Take Bet	Win $10
Decline Bet	Status quo

And it is clear what Barry's decision should be in this situation. Taking the bet dominates declining it, and Barry should take dominating options.

What has happened? It is incredibly clear that Barry should decline the bet, yet here we have an argument that he should take the bet. If you accept that the bet should be declined, then it seems to me that there are three options available.

1. Barry never knew that the musician was Beth.
2. Barry did know that the musician was Beth, but this knowledge was destroyed by the genie's offer of the bet.
3. States of the world that are known not to obtain should still be represented in decision problems, so taking the bet is not a dominating option.

The first option is basically a form of scepticism. If the take-away message from the above discussion is that Barry doesn't know the musician is Beth, we can mount a similar argument to show that he knows next to nothing.[6] And the third

[6] The idea that interest-relativity is a way of fending off scepticism is a very prominent theme in Fantl and McGrath (2009).

option would send us back into the problems about interpreting and applying decision theory that we spent the first few pages trying to get out of.

So it seems that the best solution here, or perhaps the least bad solution, is to accept that knowledge is interest-relative. Barry did know that the musician was Beth, but the genie's offer destroyed that knowledge. When Barry was unconcerned with bets at extremely long odds on whether the musician is Beth, he knows Beth is the musician. Now that he is interested in those bets, he doesn't know that.[7]

The argument here bears more than a passing resemblance to the arguments in favour of interest-relativity that are made by Hawthorne, Stanley, and Fantl and McGrath. But I think the focus on decision theory shows how we can get to interest-relativity with very weak premises.[8] In particular, the only premises I've used to derive an interest-relative conclusion are:

1. Before the genie showed up, Barry knew the musician was Beth.
2. It's rationally permissible, *in cases like Barry's*, to take dominating options.
3. It's always right to model decision problems by including what the agent knows in the 'framework'. That is, our decision tables should include what the agent knows about the payoffs in different states, and leave off any state the agent knows not to obtain.
4. It is rationally impermissible for Barry to take the genie's offered bet.

The second premise there is *much* weaker than the principles linking knowledge and action defended in previous arguments for interest-relativity. It isn't the claim that one can always act on what one knows, or that one can only act on what one knows, or that knowledge always (or only) provides reason to act. It's just the claim that in one very specific type of situation, in particular when one has to make a relatively simple bet, which affects nobody but the person making the bet, it's rationally permissible to take a dominating option. In conjunction with

[7] On the version of IRI I'm defending, Barry is free to be interested in whatever he likes. If he started wondering about whether it would be rational to take such a bet, he loses the knowledge that Beth is the musician, even if there is no genie and the bet isn't offered. The existence of the genie's offer makes the bet a practical interest; merely wondering about the genie's offer makes the bet a cognitive interest. But both kinds of interests are relevant to knowledge.

[8] As they make clear in their (2008), Hawthorne and Stanley are interested in defending relatively strong premises linking knowledge and action independently of the argument for the interest-relativity of knowledge. What I'm doing here is showing how that conclusion does not rest on anything nearly as strong as the principles they believe, and so there is plenty of space to disagree with their general principles, but accept interest-relativity. The strategy here isn't a million miles from the point noted in Fantl and McGrath (2009, 72n14) when they note that much weaker premises than the ones they endorse imply a failure of 'purism'.

the third premise, it entails that *in those kind of cases*, the fact that one knows taking the bet will lead to a better outcome suffices for making acceptance of the bet rationally permissible. It doesn't say anything about what else might or might not make acceptance rationally permissible. It doesn't say anything about what suffices for rationally permissibility in other kinds of cases, such as cases where someone else's interests are at stake, or where taking the bet might violate a deontological constraint, or any other way in which real-life choices differ from the simplest decision problems.[9] It doesn't say anything about any other kind of permissibility, e.g., moral permissibility. But it doesn't need to, because we're only in the business of proving that there is *some* interest-relativity to knowledge, and an assumption about practical rationality in some range of cases suffices to prove that.[10]

The case of Barry and Beth also bears some relationship to one of the kinds of case that have motivated contextualism about knowledge. Indeed, it has been widely noted in the literature on interest-relativity that interest-relativity can explain away many of the puzzles that motivate contextualism. And there are difficulties that face any contextualist theory (Weatherson, 2006). So I prefer an *invariantist* form of interest-relativity about knowledge. That is, my view is a form of interest-relative-invariantism, or IRI.[11]

Now everything I've said here leaves it open whether the interest-relativity of knowledge is a natural and intuitive theory, or whether it is a somewhat unhappy concession to difficulties that the case of Barry and Beth raise. I think the former is correct, and interest-relativity is fairly plausible on its own merits, but it would be consistent with my broader conclusions to say that in fact the interest-relative theory of knowledge is very implausible and counterintuitive. If we said that, we could still justify the interest-relative theory by noting that we have on our hands here a paradoxical situation, and any option will be somewhat implausible. This consideration has a bearing on how we should think about the role of intuitions

[9] I have more to say about those cases in section 2.2.

[10] Also note that I'm not taking as a premise any claim about what Barry knows after the bet is offered. A lot of work on interest-relativity has used such premises, or premises about related intuitions. This seems like a misuse of the method of cases to me. That's not because we should never use intuitions about cases, just that these cases are too hard to think that snap judgments about them are particularly reliable. In general, we can know a lot about cases by quickly reflecting on them. Similarly, we know a lot about which shelves are level and which are uneven by visual inspection, i.e., 'eyeballing'. But when different eyeballs disagree, it's time to bring in other tools. That's the approach of this paper. I don't have a story about why the various eyeballs disagree about cases like Barry's; that seems like a task best undertaken by a psychologist not a philosopher (Ichikawa, 2009).

[11] This is obviously not a full argument against contextualism; that would require a much longer paper than this.

about cases, or principles, in arguments that knowledge is interest-relative. Several critics of the view have argued that the view is counter-intuitive, or that it doesn't accord with the reactions of non-expert judges.[12] In a companion paper, "Defending Interest-Relative Invariantism", I note that those arguments usually misconstrue what the consequences of interest-relative theories of knowledge are. But even if they don't, I don't think there's any quick argument that if interest-relativity is counter-intuitive, it is false. After all, the only alternatives that seem to be open here are very counter-intuitive.

Finally, it's worth noting that if Barry is rational, he'll stop (fully) believing that the musician is Beth once the genie makes the offer. Assuming the genie allows this, it would be very natural for Barry to try to acquire more information about the singer. He might walk over to the window to see if he can see who is performing in the park. So this case leaves it open whether the interest-relativity of knowledge can be explained fully by the interest-relativity of belief. I used to think it could be; I no longer think that. To see why this is so, it's worth rehearsing how the interest-relative theory of belief runs.

5.2 Playing Games with a Lockean

I'm going to raise problems for Lockeans, and for defenders of regularity in general, by discussing a simple game. 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 it is different to what I used to think, and perhaps to what some other people think too, so I'll set it out slowly.[13]

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 able.

	S_1	S_2
A_1	4	0
A_2	1	1

[12] See, for instance, Blome-Tillmann (2009), or Feltz and Zarpentine (2010).

[13] 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.

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 that we discussed a couple of paragraphs ago. The relevant states of the world are choices of Column. But that's fine; we didn't say in setting out the decision problem what constituted the states S_1 and S_2. And note that we solved the problem without explicitly saying anything about probabilities. What we added was some information about Column's payouts, and the fact that Column is rational. From there we deduced something about Column's play, namely that she would play S_2. And from that we concluded what Row should 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. (We'll come back to whether the 'mostly' qualification is needed below.) 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, we can't say much about a decision problem with unspecified probabilities. But we can and do say a lot about games where the setup of the game doesn't specify the probabilities, because we can solve for them given the other information we have.

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. Arguably the core solution concept in game theory is the Nash equilibrium. As you probably know, 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 puzzles. 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.[14] 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 Red-Green. I've said something different about this game in earlier work (Weatherson, 2012a). But I now think that to understand what's going on, we need to think about mixed strategies where one element of the mixture has probability one.

Informally, in this game A and B must each play either a green or red card. I will capitalise A's moves, i.e., A can play GREEN or RED, and italicise B's moves, i.e., B can play *green* or *red*. If two green cards, or one green card and one red card are played, each player gets \$1. If two red cards 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.

	green	*red*
GREEN	1, 1	1, 1
RED	1, 1	0, 0

[14] Actually, I guess 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.

When I write game tables like this, and I think this is the usual way game tables are to be interpreted (Weatherson, 2012b), 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. 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. We'll relax this later, but it will help the discussion a lot I think 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 (White, 2005). I don't, but it does seem like it might *often* hold. 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 GREEN should be the same as the probability of *B* playing *green*, 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 GREEN, is 1. Indeed, the known value of GREEN is 1. If the probability of *green* is x, then the expected value of RED is x. So if $x < 1$, and Row is rational, she'll definitely play GREEN. But that's inconsistent with the claim that $x < 1$, since that means that it isn't definite that Row will play GREEN.

So we can conclude that $x = 1$. Does that mean we can know that Row will play GREEN? 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 *green*. Since any rational person can conclude this, Row can conclude it. So Row knows that she'll get 1 whether she plays GREEN or RED. But then she should be indifferent between playing GREEN and RED. And if we know she's indifferent between playing GREEN and RED, 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 GREEN.

I think the arguments of the last two paragraphs are sound. We'll turn to an objection presently, but let's note how bizarre is the conclusion we've reached. One argument has shown that it could not be more probable that Row will play GREEN. A second argument has shown that we can't know that Row will play GREEN. It reminds me of examples involving blindspots (Sorensen, 1988). Consider this case:

(B) Brian does not know (B).

That's true, right? Assume it's false, so I do know (B). Knowledge is factive, so (B) is true. But that contradicts the assumption that it's false. So it's true. But I obviously don't know that it's true; that's what this very true proposition says.

Now I'm not going to rest anything on this case, because there are so many tricky things one can say about blindspots, and about the paradoxes generally. It does suggest that there are other finite cases where one can properly have maximal credence in a true proposition without knowledge.[15] And, assuming that we shouldn't believe things we know we don't know, that means we can have maximal credence in things we don't believe. All I want to point out is that this phenomena of maximal credence without knowledge, and presumably without full belief, isn't a quirky feature of self-reference, or of games, or of puzzles about infinity; it comes up in a wide range of cases.

For the rest of this section 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 GREEN, we could show that *A* had no reason to play GREEN. But actually what we showed was that the expected utility of playing GREEN would be the same as playing RED. Perhaps *A* has a reason to play GREEN, namely that GREEN weakly dominates RED. After all, there's one possibility on the table where GREEN does better than RED, 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 GREEN 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

[15] As an aside, the existence of these cases is why I get so irritated when epistemologists try to theorise about 'Gettier Cases' as a class. What does (B) have in common with inferences from a justified false belief, or with otherwise sound reasoning that is ever so close to issuing in a false conclusion due to relatively bad luck? As far as I can tell, the class of justified true beliefs that aren't knowledge is a disjunctive mess, and this should matter for thinking about the nature of knowledge. For further examples, see Williamson (2012).

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 GREEN/*green*, 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 *blue*, and if *B* plays *blue*, *A* will receive 2 if she plays RED, and -1 if she plays GREEN. Well hold on, you might think, didn't I say that *green* and *red* 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, GREEN does better than RED.

We can put this as a dilemma. Assume, for *reductio*, that GREEN/*green* is the only rational play. Then if we restrict our attention to possibilities that are epistemically open to *A*, then GREEN does just as well as RED; 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 RED. Since *A* maximises expected utility, that means *A* must have credence 1 that *B* will play *green*. 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 *green*, it isn't obvious that playing RED is rational. After all, relative to the space of epistemic possibilities, GREEN weakly dominates RED. 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 RED. 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 GREEN, like whether the coin will ever land tails, is a case the Lockean cannot get right, no matter where they set the threshold for belief; our credence is above the threshold, but we don't believe.

So I think this case is a real problem for a Lockean view about the relationship between credence and belief. If A is rational, she can have credence 1 that B will play *green*, but won't believe that B will play *green*. But now you might worry that my own account of the relationship between belief and credence is in just as much trouble. After all, I said that to believe p is, roughly, to have the same attitudes towards all salient questions as you have conditional on p. And it's hard to identify a question that rational A would answer differently upon conditionalising on the proposition that B plays *green*.

I think what went wrong in my earlier view was that I'd too quickly equated updating with conditionalisation. The two can come apart. Here's an example from Gillies (2010) that makes the point well.

> 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)

As Gillies goes on to point out, this isn't really a problem for the Ramsey test view of conditionals.

> 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.

For example, in the game I've been discussing, we should believe that rational *B* might play *red*. Indeed, the probability of that is, I think, 1. And whether or not *B* might play red is highly salient; it matters to the probability of whether **A** will play GREEN or RED. Conditionalising on something that has probability 1, such as that *B* will play *green*, can hardly change that probability. But updating on the proposition that *B* will play *green* can make a difference. We can see that by simply noting that the conditional *If B plays green, she might play red* is incoherent.

So I conclude that a theory of belief like mine can handle the puzzle this game poses, as long as it distinguishes between conditionalising and updating, in just the way Gillies suggests. To believe that *p* is to be disposed to not change any attitude towards a salient question on updating that *p*. (Plus some bells and whistles to deal with propositions that are not relevant to salient questions. We'll return to them below.) Updating often goes by conditionalisation, so we can often say that belief means having attitudes that match unconditionally and conditionally on *p*. But not all updating works that way, and the theory of belief needs to acknowledge this.

CHAPTER 6

KNOWLEDGE AND JUSTIFICATION

6.1 Interest-Relative Defeaters

As I said at the top, I've changed my view from Doxastic IRI to Non-Doxastic IRI. The change of heart is occasioned by cases like the following, where the agent is mistaken, and hence ignorant, about the odds at which she is offered a bet on p. In fact the odds are much longer than she thinks. Relative to what she stands to win, the stakes are too high.

6.1.1 The Coraline Example

The problem for Doxastic IRI arises because of cases like that of Coraline. Here's what we're going to stipulate about Coraline.

- She knows that p and q are independent, so her credence in any conjunction where one conjunct is a member of $\{p, \neg p\}$ and the other is a member of $\{q, \neg q\}$ will be the product of her credences in the conjuncts.
- Her credence in p is 0.99, just as the evidence supports.
- Her credence in q is also 0.99. This is unfortunate, since the rational credence in q given her evidence is 0.01.
- The only relevant question for her which is sensitive to p is whether to take or decline a bet with the following payoff structure.[1] (Assume that the marginal utility of money is close enough to constant that expected dollar returns correlate more or less precisely with expected utility returns.)

[1] I'm more interested in the abstract structure of the case than in whether any real-life situation is modelled by just this structure. But it might be worth noting the rough kind of situation where this kind of situation can arise. So let's say Coraline has a particular bank account that is uninsured, but which currently paying 10% interest, and she is deciding whether to deposit another $1000 in it. Then p is the proposition that the bank will not collapse, and she'll get her money back, and q is the proposition that the interest will stay at 10%. To make the model exact, we have to also assume that if the interest rate on her account doesn't stay at 10%, it falls to 0.1%. And we have to assume that the interest rate and the bank's collapse are probabilistically independent. Neither of these are at all realistic, but a realistic case would simply be more complicated, and the complications would obscure the philosophically interesting point.

	$p \wedge q$	$p \wedge \neg q$	$\neg p$
Take bet	$100	$1	−$1000
Decline bet	0	0	0

As can be easily computed, the expected utility of taking the bet given her credences is positive, it is just over $89. And Coraline takes the bet. She doesn't compute the expected utility, but she is sensitive to it.[2] That is, had the expected utility given her credences been close to 0, she would have not acted until she made a computation. But from her perspective this looks like basically a free $100, so she takes it. Happily, this all turns out well enough, since p is true. But it was a dumb thing to do. The expected utility of taking the bet given her evidence is negative, it is a little under -$8. So she isn't warranted, given her evidence, in taking the bet.

6.1.2 What Coraline Knows and What She Believes

Assume, for *reductio*, that Coraline knows that p. Then the choice she faces looks like this.

	q	$\neg q$
Take bet	$100	$1
Decline bet	0	0

Since taking the bet dominates declining the bet, she should take the bet if this is the correct representation of her situation. She shouldn't take the bet, so by *modus tollens*, that can't be the correct representation of her situation. If she knew p, that would be the correct representation of her situation. So, again by *modus tollens*, she doesn't know p.

Now let's consider three possible explanations of why she doesn't know that p.

1. She doesn't have enough evidence to know that p, independent of the practical stakes.
2. In virtue of the practical stakes, she doesn't believe that p;
3. In virtue of the practical stakes, she doesn't justifiably believe that p, although she does actually believe it.
4. In virtue of the practical stakes, she doesn't know that p, although she does justifiably believe it.

[2] If she did compute the expected utility, then one of the things that would be salient for her is the expected utility of the bet. And the expected utility of the bet is different to its expected utility given p. So if that expected utility is salient, she doesn't believe p. And it's going to be important to what follows that she *does* believe p.

I think option 1 is implausibly sceptical, at least if applied to all cases like Coraline's. I've said that the probability of p is 0.99, but it should be clear that all that matters to generating a case like this is that p is not completely certain. Unless knowledge requires certainty, we'll be able to generate Coraline-like cases where there is sufficient evidence for knowledge. So that's ruled out.

Option 2 is basically what the Doxastic IRI theorist has to say. If Coraline has enough evidence to know p, but doesn't know p due to practical stakes, then the Doxastic IRI theorist is committed to saying that the practical stakes block *belief* in p. That's the Doxastic IRI position; stakes matter to knowledge because they matter to belief.

But that's also an implausible description of Coraline's situation. She is very confident that p. Her confidence is grounded in the evidence in the right way. She is insensitive in her actual deliberations to the difference between her evidence for p and evidence that guarantees p. She would become sensitive to that difference if someone offered her a bet that she knew was a 1000-to-1 bet on p, but she doesn't know that's what is on offer. In short, there is no difference between her unconditional attitudes, and her attitudes conditional on p, when it comes to any live question. That's enough, I think, for belief. So she believes that p. And that's bad news for the Doxastic IRI theorist; since it means here that stakes matter to knowledge without mattering to belief. I conclude, reluctantly, that Doxastic IRI is false.

6.1.3 Stakes as Defeaters

That still leaves two options remaining, what I've called options 3 and 4 above. Option 3, if suitably generalised, says that knowledge is practically sensitive because the justification condition on belief is practically sensitive. Option 4 says that practical considerations impact knowledge directly. As I read them, Jeremy Fantl and Matthew McGrath defend a version of Option 3. In the next and last subsection, I'll argue against that position. But first I want to sketch what a position like option 4 would look like.

Knowledge, unlike justification, requires a certain amount of internal coherence amongst mental states. Consider the following story from David Lewis:

> I speak from experience as the repository of a mildly inconsistent corpus. I used to think that Nassau Street ran roughly east-west; that the railroad nearby ran roughly north-south; and that the two were roughly parallel. (Lewis, 1982, 436)

I think in that case that Lewis doesn't know that Nassau Street runs roughly east-west. (From here on, call the proposition that Nassau Street runs roughly east-west N.) If his belief that it does was acquired and sustained in a suitably reliable way, then he may well have a justified belief that N. But the lack of coherence with the rest of his cognitive system, I think, defeats any claim to knowledge he has.

Coherence isn't just a requirement on belief; other states can cohere or be incoherent. Assume Lewis corrects the incoherence in his beliefs, and drops the belief that Nassau Street the railway are roughly parallel. Still, if Lewis believed that N, preferred doing φ to doing ψ conditional on N, but actually preferred doing ψ to doing φ, his cognitive system would also be in tension. That tension could, I think, be sufficient to defeat a claim to know that N.

And it isn't just a requirement on actual states; it can be a requirement on rational states. Assume Lewis believed that N, preferred doing φ to doing ψ conditional on N, and preferred doing φ to doing ψ, but should have preferred doing ψ to doing φ given his interests. Then I think the fact that the last preference is irrational, plus the fact that were it corrected there would be incoherence in his cognitive states defeats the claim to know that N.

A concrete example of this helps make clear why such a view is attractive, and why it faces difficulties. Assume there is a bet that wins \$2 if N, and loses \$10 if not. Let φ be taking that bet, and ψ be declining it. Assume Lewis shouldn't take that bet; he doesnt have enough evidence to do so. Then he clearly doesn't know that N. If he knew that N, φ would dominate ψ, and hence be rational. But it isn't, so N isn't known. And that's true whether Lewis's preferences between φ and ψ are rational or irrational.

Attentive readers will see where this is going. Change the bet so it wins a penny if N, and loses \$1,000 if not. Unless Lewis's evidence that N is incredibly strong, he shouldn't take the bet. So, by the same reasoning, he doesn't know that N. And we're back saying that knowledge requires incredibly strong evidence. The solution, I say, is to put a pragmatic restriction on the kinds of incoherence that matter to knowledge. Incoherence with respect to irrelevant questions, such as whether to bet on N at extremely long odds, doesn't matter for knowledge. Incoherence (or coherence obtained only through irrationality) does. The reason, I think, that Non-Doxastic IRI is true is that this coherence-based defeater is sensitive to practical interests.

The string of cases about Lewis and N has ended up close to the Coraline example. We already concluded that Coraline didn't know p. Now we have a story about why she doesn't know that p. Her belief that p doesn't cohere sufficiently well with what she should believe, namely that it would be wrong to take the bet. If all that is correct, just one question remains: does this coherence-based defeater also defeat Coraline's claim to have a justified belief that p? I say it does not, for three reasons.

First, her attitude towards p tracks the evidence perfectly. She is making no mistakes with respect to p. She is making a mistake with respect to q, but not with respect to p. So her attitude towards p, i.e. belief, is justified.

Second, talking about beliefs and talking about credences are simply two ways of modelling the very same things, namely minds. If the agent both has a credence 0.99 in p, and believes that p, these are not two different states. Rather, there is one state of the agent, and two different ways of modelling it. So it is implausible to apply different valuations to the state depending on which modelling tools we choose to use. That is, it's implausible to say that while we're modelling the agent with credences, the state is justified, but when we change tools, and start using beliefs, the state is unjustified. Given this outlook on beliefs and credences, it is natural to say that her belief is justified. Natural, but not compulsory, for reasons Jeremy Fantl pointed out to me.[3] We don't want a metaphysics on which persons and philosophers are separate entities. Yet we can say that someone is a good person but a bad philosopher. Normative statuses can differ depending on which property of a thing we are considering. That suggests it is at least coherent to say that one and the same state is a good credence but a bad belief. But while this may be coherent, I don't think it is well motivated, and it is natural to have the evaluations go together.

Third, we don't *need* to say that Coraline's belief in p is unjustified in order to preserve other nice theories, in the way that we do need to say that she doesn't know p in order to preserve a nice account of how we understand decision tables. It's this last point that I think Fantl and McGrath, who say that the belief is unjustified, would reject. So let's conclude with a look at their arguments.

6.1.4 Fantl and McGrath on Interest-Relativity

Fantl and McGrath's argue for the principle (JJ), which entails that Coraline is not justified in believing p.

(JJ) If you are justified in believing that p, then p is warranted enough to justify you in φ-ing, for any φ. (Fantl and McGrath, 2009, 99)

In practice, what this means is that there can't be a salient p, φ such that:

- The agent is justified in believing p;

[3] The following isn't Fantl's example, but I think it makes much the same point as the examples he suggested.

- The agent is not warranted in doing φ; but
- If the agent had more evidence for p, and nothing else, the agent would be warranted in doing φ.

That is, once you've got enough evidence, or warrant, for justified belief in p, then you've got enough evidence for p as matters for any decision you face. This seems intuitive, and Fantl and McGrath back up its intuitiveness with some nicely drawn examples. But I think it is false, and the Coraline example shows it is false. Coraline isn't justified in taking the bet, and is justified in believing p, but more evidence for p would suffice for taking the bet. So Coraline's case shows that (JJ) is false. But there are a number of possible objections to that position. I'll spend the rest of this section, and this paper, going over them.[4]

Objection: The following argument shows that Coraline is not in fact justified in believing that p.

1. p entails that Coraline should take the bet, and Coraline knows this.
2. If p entails something, and Coraline knows this, and she justifiably believes p, she is in a position to justifiably believe the thing entailed.
3. Coraline is not in a position to justifiably believe that she should take the bet.

C. So, Coraline does not justifiably believe that p

Reply: The problem here is that premise 1 is false. What's true is that p entails that Coraline will be better off taking the bet than declining it. But it doesn't follow that she should take the bet. Indeed, it isn't actually true that she should take the bet, even though p is actually true. Not just is the entailment claim false, the world of the example is a counterinstance to it.

It might be controversial to use this very case to reject premise 1. But the falsity of premise 1 should be clear on independent grounds. What p entails is that Coraline will be best off by taking the bet. But there are lots of things that will make me better off that I shouldn't do. Imagine I'm standing by a roulette wheel, and the thing that will make me best off is betting heavily on the number than will actually come up. It doesn't follow that I should do that. Indeed, I should not do it. I shouldn't place any bets at all, since all the bets have a highly negative expected return.

[4] Thanks here to a long blog comments thread with Jeremy Fantl and Matthew McGrath for making me formulate these points much more carefully. The original thread is at http://tar.weatherson.org/2010/03/31/do-justified-beliefs-justify-action/.

In short, all p entails is that taking the bet will have the best consequences. Only a very crude kind of consequentialism would identify what I should do with what will have the best returns, and that crude consequentialism isn't true. So p doesn't entail that Coraline should take the bet. So premise 1 is false.

Objection: Even though p doesn't *entail* that Coraline should take the bet, it does provide inductive support for her taking the bet. So if she could justifiably believe p, she could justifiably (but non-deductively) infer that she should take the bet. Since she can't justifiably infer that, she isn't justified in taking the bet.

Reply: The inductive inference here looks weak. One way to make the inductive inference work would be to deduce from p that taking the bet will have the best outcomes, and infer from that that the bet should be taken. But the last step doesn't even look like a reliable ampliative inference. The usual situation is that the best outcome comes from taking an *ex ante* unjustifiable risk.

It may seem better to use p combined with the fact that conditional on p, taking the bet has the highest *expected* utility. But actually that's still not much of a reason to take the bet. Think again about cases, completely normal cases, where the action with the best outcome is an *ex ante* unjustifiable risk. Call that action φ, and let $B\varphi$ be the proposition that φ has the best outcome. Then $B\varphi$ is true, and conditional on $B\varphi$, φ has an excellent expected return. But doing φ is still running a dumb risk. Since these kinds of cases are normal, it seems it will very often be the case that this form of inference leads from truth to falsity. So it's not a reliable inductive inference.

Objection: In the example, Coraline isn't just in a position to justifiably believe p, she is in a position to *know* that she justifiably believes it. And from the fact that she justifiably believes p, and the fact that if p, then taking the bet has the best option, she can infer that she should take the bet.

Reply: It's possible at this point that we get to a dialectical impasse. I think this inference is non-deductive, because I think the example we're discussing here is one where the premises are true and the conclusion false. Presumably someone who doesn't like the example will think that it is a good deductive inference.

Having said that, the more complicated example at the end of Weatherson (2005a) was designed to raise the same problem without the consequence that if p is true, the bet is sure to return a positive amount. In that example, conditionalising on p means the bet has a positive expected return, but still possibly a negative return. But in that case (JJ) still failed. If accepting there are cases where an agent justifiably believes p, and hence justifiably believes taking the bet will return the best outcome, and knows all this, but still can't rationally bet on p is too much to accept, that more complicated example might be more persuasive. Otherwise, I concede that someone who believes (JJ) and thinks rational agents can use it in their reasoning will not think that a particular case is a counterexample to (JJ).

Objection: If Coraline were ideal, then she wouldn't believe p. That's because if she were ideal, she would have a lower credence in q, and if that were the case, her credence in p would have to be much higher (close to 0.999) in order to count as a belief. So her belief is not justified.

Reply: The premise here, that if Coraline were ideal she would not believe that p, is true. The conclusion, that she is not justified in believing p, does not follow. It's always a mistake to *identify* what should be done with what is done in ideal circumstances. This is something that has long been known in economics. The *locus classicus* of the view that this is a mistake is Lipsey and Lancaster (1956-1957). A similar point has been made in ethics in papers such as Watson (1977) and Kennett and Smith (1996a,b). And it has been extended to epistemology by Williamson (1998).

All of these discussions have a common structure. It is first observed that the ideal is both F and G. It is then stipulated that whatever happens, the thing being created (either a social system, an action, or a cognitive state) will not be F. It is then argued that given the stipulation, the thing being created should not be G. That is not just the claim that we shouldn't *aim* to make the thing be G. It is, rather, that in many cases being G is not the best way to be, given that F-ness will not be achieved. Lipsey and Lancaster argue that (in an admittedly idealised model) that it is actually quite unusual for G to be best given that the system being created will not be F.

It's not too hard to come up with examples that fit this structure. Following (Williamson, 2000, 209), we might note that I'm justified in believing that there are no ideal cognitive agents, although were I ideal I would not believe this. Or imagine a student taking a ten question mathematics exam who has no idea how to answer the last question. She knows an ideal student would correctly answer an even number of questions, but that's no reason for her to throw out her good answer to question nine. In general, once we have stipulated one departure from the ideal, there's no reason to assign any positive status to other similarities to the idea. In particular, given that Coraline has an irrational view towards q, she won't perfectly match up with the ideal, so there's no reason it's good to agree with the ideal in other respects, such as not believing p.

Stepping back a bit, there's a reason the interest-relative theory says that the ideal and justification come apart right here. On the interest-relative theory, as on any pragmatic theory of mental states, the *identification* of mental states is a somewhat holistic matter. Something is a belief in virtue of its position in a much broader network. But the *evaluation* of belief is (relatively) atomistic. That's why Coraline is justified in believing p, although if she were wiser she would not believe it. If she were wiser, i.e., if she had the right attitude towards q, the very same credence in p would not count as a belief. Whether her state counts as a belief, that is, depends on wide-ranging features of her cognitive system. But whether the state is justified depends on more local factors, and in local respects she is doing everything right.

Objection: If Coraline is justified in believing p, then Coraline can use p as a premise in practical reasoning. If Coraline can use p as a premise in practical reasoning, and p is true, and her belief in p is not Gettiered, then she knows p. By hypothesis, her belief is true, and her belief is not Gettiered. So she should know p. But she doesn't know p. So by several steps of modus tollens, she isn't justified in believing p.[5]

Reply: This objection this one turns on an equivocation over the neologism 'Gettiered'. Some epistemologists use this to simply mean that a belief is justified and true without constituting knowledge. By that standard, the third sentence is false. Or, at least, we haven't been given any reason to think that it is true. Given everything else that's said, the third sentence is a raw assertion that Coraline knows that p, and I don't think we should accept that.

The other way epistemologists sometimes use the term is to pick out justified true beliefs that fail to be knowledge for the reasons that the beliefs in the original examples from Gettier (1963) fail to be knowledge. That is, it picks out a property that beliefs have when they are derived from a false lemma, or whatever similar property is held to be doing the work in the original Gettier examples. Now on this reading, Coraline's belief that p is not Gettiered. But it doesn't follow that it is known. There's no reason, once we've given up on the JTB theory of knowledge, to think that whatever goes wrong in Gettier's examples is the *only* way for a justified true belief to fall short of knowledge. It could be that there's a practical defeater, as in this case. So the second sentence of the objection is false, and the objection again fails.

Once we have an expansive theory of defeaters, as I've adopted here, it becomes problematic to describe the case in the language Fantl and McGrath use. They focus a lot on whether agents like Coraline have 'knowledge-level justification' for p, which is defined as "justification strong enough so that shortcomings in your strength of justification stand in the way of your knowing". (Fantl and

[5] Compare the 'subtraction argument' on page 99 of Fantl and McGrath (2009).

McGrath, 2009, 97). An important part of their argument is that an agent is justified in believing p iff they have knowledge-level justification for p. I haven't addressed this argument, so I'm not really addressing the case on their terms.

Well, does Coraline have knowledge-level justification for p? I'm not sure, because I'm not sure I grasp this concept. Compare the agent in Harman's dead dictator case (Harman, 1973, 75). Does she have knowledge-level justification that the dictator is dead? In one sense yes; it is the existence of misleading news sources that stops her knowing. In another sense no; she doesn't know, but if she had better evidence (e.g., seeing the death happen) she would know. I want to say the same thing about Coraline, and that makes it hard to translate the Coraline case into Fantl and McGrath's terminology.

6.2 *Measurement, Justification and Knowledge*

Williamson's core example involves detecting the angle of a pointer on a wheel by eyesight. For various reasons, I find it easier to think about a slightly different example: measuring a quantity using a digital measurement device. This change has some costs relative to Williamson's version – for one thing, if we are measuring a quantity it might seem that the margin of error is related to the quantity measured. If I eyeball how many stories tall a building is, my margin of error is 0 if the building is 1-2 stories tall, and over 10 if the building is as tall as the World Trade Center. But this problem is not as pressing for digital devices, which are often very *unreliable* for small quantities. And, at least relative to my preferences, the familiarity of quantities makes up for the loss of symmetry properties involved in angular measurement.[6]

To make things explicit, I'll imagine the agent S is using a digital scale. The scale has a **margin of error** m. That means that if the reading, i.e., the **apparent mass** is a, then the agent is justified in believing that the mass is in $[a-m, a+m]$. We will assume that a and m are luminous; i.e., the agent knows their values, and knows she knows them, and so on. This is a relatively harmless idealisation for a; it is pretty clear what a digital scale reads.[7] It is a somewhat less plausible assumption for m. But we'll assume that S has been very diligent about calibrating

[6] There's one other change that might make a larger difference. When we use a digital device, there's a very clear separation between what the input is, and what we do with that input. That kind of factorisation is not nearly as easy when we are eyeballing something, and may well be impossible. But I think it makes the discussion smoother to have a case where we can easily separate the input from the processing.

[7] This isn't always true. If a scale flickers between reading 832g and 833g, it takes a bit of skill to determine what *the reading* is. But we'll assume it is clear in this case. On an analogue scale, the luminosity assumption is rather implausible, since it is possible to eyeball with less than perfect accuracy how far between one marker and the next the pointer is.

her scale, and that the calibration has been recently and skillfully carried out, so in practice m can be assessed very accurately.

Note that when I say the scale has a margin of error, this is a tensed claim. The margin will change over time, and with a change of environment and so on. Many scales will come with a designed margin of error, which may be printed on the scale or even on the display. I'm *not* assuming that m is that margin. In fact, it will be easiest in what follows if we assume that m is much much larger than this printed value. This could be because the scale is old, or because it is being used in a noisy environment. What we are assuming is that S has very carefully assessed the accuracy of the scale in the environment she is using it, and correctly concluded that when it reads a, she is justified in believing that the mass of the object on the scale is in $[a-m, a+m]$.

We'll make three further assumptions about m that strike me as plausible, but which may I guess be challenged. I need to be a bit careful with terminology to set out the first one. I'll use V and v as variables that both pick out the **true value** of the mass. The difference is that v picks it out rigidly, while V picks out the value of the mass in any world under consideration. Think of V as shorthand for *the mass of the object* and v as shorthand for *the actual mass of the object*. (More carefully, V is a *random* variable, while v is a standard, rigid, variable.) Our first assumption then is that m is also related to what the agent can know. In particular, we'll assume that if the reading a equals v, then the agent can know that $V \in [a-m, a+m]$, and can't know anything stronger than that. That is, the margin of error for justification equals, in the best case, the margin of error for knowledge. The second is that the scale has a readout that is finer than m. This is usually the case; the last digit on a digital scale is often not significant. The final assumption is that it is metaphysically possible that the scale has an error on an occasion that is greater than m. This is a kind of fallibilism assumption – saying that the margin of error is m does not mean there is anything incoherent about talking about cases where the error on an occasion is greater than m.

This error term will do a lot of work in what follows, so I'll use e to be the **error** of the measurement, i.e., $|a-v|$. For ease of exposition, I'll assume that $a \geq v$, i.e., that any error is on the high side. But this is entirely dispensible, and just lets me drop some disjunctions later on.

Now we are in a position to state Williamson's argument. Assume that on a particular occasion, $0 < e < m$. Perhaps $v = 830, m = 10$ and $a = 832$, so $e = 2$. Williamson appears to make the following two assumptions.[8]

[8] I'm not actually sure whether Williamson *makes* the first, or thinks it is the kind of thing anyone who thinks justification is prior to knowledge should make.

1. The agent is justified in believing what they would know if appearances matched reality, i.e., if V equalled a.
2. The agent cannot come to know something about V on the basis of a suboptimal measurement that they could not also know on the basis of an optimal measurement.

I'm assuming here that the optimal measurement displays the correct mass. I don't assume the actual measurement is *wrong*. That would require saying something implausible about the semantic content of the display. It's not obvious that the display has a content that could be true or false, and if it does have such a content it might be true. (For instance, the content might be that the object on the scale has a mass near to a, or that with a high probability it has a mass near to a, and both of those things are true. Or it might be that the mass of the object on the scale is within the printed margin of error of the scale. Even if that's false in the case we're imagining, it could be true without $a = v$.) But the optimal measurement would be to have $a = v$, and in this sense the measurement is suboptimal.

The argument then is pretty quick. From the first assumption, we get that the agent is justified in believing that $V \in [a-m, a+m]$. Assume then that the agent forms this justified belief. This belief is incompatible with $V \in [v-m, a-m)$. But if a equalled v, then the agent wouldn't be in a position to rule out that $V \in [v-m, a-m)$. So by premise 2 she can't knowledgeably rule it out on the basis of a mismeasurement. So her belief that $V \geq a-m$ cannot be knowledge. So this justified true belief is not knowledge.

If you prefer doing this with numbers, here's the way the example works using the numbers above. The mass of the object is 830. So if the reading was correct, the agent would know just that the mass is between 820 and 840. The reading is 832. So she's justified in believing, and we'll assume she does believe, that the mass is between 822 and 842. That belief is incompatible with the mass being 821. But by premise 2 she can't know the mass is greater than 821. So the belief doesn't amount to knowledge, despite being justified and, crucially, true. After all, 830 is between 822 and 842, so her belief that the mass is in this range is true. So simple reflections on the workings on measuring devices let us generate cases of justified true beliefs that are not knowledge.

I'll end this section with a couple of objections and replies.

Objection: The argument that the agent can't know that $V \in [a-m, a+m]$ is also an argument that the argument can't justifiably believe that $V \in [a-m, a+m]$. After all, why should it be possible to get justification from a suboptimal measurement when it isn't possible to get the same justification from an optimal measurement?

Reply: It is possible to have justification to believe an outright falsehood. It is widely believed that you can have justification even when none of your evidential sources are even approximately accurate (Cohen, 1984). And even most reliabilists will say that you can have false justified beliefs if you use a belief forming method that is normally reliable, but which badly misfires on this occasion. In such cases we clearly get justification to believe something from a mismeasurement that we wouldn't get from a correct measurement. So the objection is based on a mistaken view of justification.

Objection: Premise 2 fails in cases using random sampling. Here's an illustration. An experimenter wants to know what percentage of Fs are G. She designs a survey to ask people whether they are G. The survey is well designed; everyone gives the correct answer about themselves. And she designs a process for randomly sampling the Fs to get a good random selection of 500. It's an excellent process; every F had an equal chance of being selected, and the sample fairly represents the different demographically significant subgroups of the Fs. But by the normal processes of random variation, her group contains slightly more Gs than the average. In her survey, 28% of people said (truly!) that they were G, while only 26% of Fs are Gs. Assuming a margin of error in such a study of 4%, it seems plausible to say she knows that between 25 and 32% of Fs are Gs. But that's not something she could have known the survey had come back correctly reporting that 26% of Fs are Gs.

Reply: I think the core problem with this argument comes in the last sentence. A random survey isn't, in the first instance, a measurement of a population. It's a measurement of those surveyed, from which we draw extrapolations about the population. In that sense, the only *measurement* in the imagined example was as good as it could be; 28% of surveyed people are in fact G. So the survey was correct, and it is fine to conclude that we can in fact know that between 24 and 32 percent of Fs are Gs.

There are independent reasons for thinking this is the right way to talk about the case. If a genuine measuring device, like a scale, is off by a small amount, we regard that as a reason for tinkering with the device, and trying to make it more accurate. That's one respect in which the measurement is suboptimal, even if it is correct within the margin of error. This reason to tinker with the scale is a reason that often will be outweighed. Perhaps it is technologically infeasible to make the machine more accurate. More commonly, the only way to guarantee greater accuracy would be more cost and hassle than it is worth. But it remains a reason. The fact that this experiment came out with a deviation between the sample and the population is *not* a reason to think that it could have been run in a better way, or that there is some reason to improve the survey. That's just how random sampling goes. If it were a genuine measurement of the population, the deviation between the 'measurement' and what is being measured would be a reason to do things differently. There isn't any such reason, so the sample is not truly a measurement.

So I don't think this objection works, and I think the general principle that you can't get extra knowledge from a suboptimal measurement is right. But note also that we don't need this general principle to suggest that there will be cases of justified true belief without knowledge in the cases of measurement. Consider a special case where e is just less than m. For concreteness, say $a = v + 0.95m$, so $e = 0.95m$. Now assume that whatever is justifiedly truly believed in this case is known, so S knows that $V \in [a - m, a + m]$. That is, S knows that $V \in [v - 0.05m, a + m]$.

We don't need any principles about measurement to show this is false; safety considerations will suffice. Williamson (2000) says that a belief that p is safe only if p is true in all nearby worlds. But given how close v is to the edge of the range $[v - 0.05m, a + m]$, the belief that v is in this range clearly isn't safe, so isn't knowledge. Rival conceptions of safety don't help much more than this. The most prominent of these, suggested by Sainsbury (1996), says that a belief is safe only if the method that produced it doesn't produce a false belief in any nearby world. But if the scale was off by $0.95m$, it could have been off by $1.05m$, so that condition fails too.

I don't want the last two paragraphs to leave too concessive an impression. I think the objection fails because it relies on a misconception of the notion of measurement. But I think that even if the objection works, we can get a safety based argument that some measurement cases will produce justified true beliefs without knowledge. And that will matter for the argument of the next two sections.

6.3 The Class of Gettier Cases is Disjunctive

There's an unfortunate terminological confusion surrounding gaps between knowledge and justification. Some philosophers use the phrase 'Gettier case' to describe any case of a justified true belief that isn't knowledge. Others use it to describe just cases that look like the cases in Gettier (1963), i.e., cases of true belief derived from justified false belief. I don't particularly have strong views on whether either of these uses is *better*, but I do think it is important to keep them apart.

I'll illustrate the importance of this by discussing a recent argument due to Jeremy Fantl and Matthew McGrath (Fantl and McGrath, 2009, Ch. 4). I've previously discussed this argument (Weatherson, 2011), but I don't think I quite got to the heart of why I don't like the kind of reasoning they are using.

The argument concerns an agent, call her T, who has the following unfortunate combination of features. She is very confident that p. And with good reason; her evidence strongly supports p. For normal reasoning, she takes p for granted. That is, she doesn't distinguish between φ is best given p, and that φ is simply best. And that's right too, given the strong evidence that p. But she's not crazy. Were she to think that she was facing a bet on extreme odds concerning p, she would cease taking p for granted, and revert to trying to maximise expected value given the high probability that p. But she doesn't think any such bet is salient, so her disposition to retreat from p to *Probably p* has not been triggered. So far, all is going well. I'm inclined to say that this is enough to say that T justifiedly believes that p. She believes that p in virtue of the fact that she takes p for granted in actual reasoning.[9] She's disposed to stop doing so in some circumstances, but until that disposition is triggered, she has the belief. And this is the right way to act given her evidence, so her belief is justified. So far, so good.

Unfortunately, T really does face a bet on long odds about p. She knows she has to choose between φ and ψ. And she knows that φ will produce the better outcome iff p. But she thinks the amount she'll gain by choosing ψ if $\neg p$ is roughly the same as the amount she'll gain by choosing φ if p. That's wrong, and her evidence clearly shows it is wrong. If p is false, then φ will be *much* worse than ψ. In fact, the potential loss here is so great that ψ has the greater expected value given the correct evidential probability of p. I think that means she doesn't know that p. Someone who knows that p can ignore $\neg p$ possibilities in practical reasoning. And someone who could ignore $\neg p$ possibilities in practical reasoning would choose φ over ψ, since it is better if p. But T isn't in a position to make that choice, so she doesn't know that p.

(I've said here that T is wrong about the costs of choosing φ if p, and her evidence shows she is wrong. In fact I think she doesn't know p if either of those conditions obtain. But here I only want to use the weaker claim that she doesn't know p if both conditions obtain.)

[9] There are some circumlocutions here because I'm being careful to be sensitive to the points raised in Ross and Schroeder (fort) about the relationship between belief and reasoning. I think there's less distance between the view they put forward and the view I defended in Weatherson (2005a) than they do, but this is a subtle matter, and for this paper's purposes I want to go along with Ross and Schroeder's picture of belief.

Fantl and McGrath agree about the knowledge claim, but disagree about the justified belief claim. They argue as follows (this is my version of the 'Subtraction Argument' from page 97 of their book).

1. T is justfied in choosing φ iff she knows that p.
2. Whether T's belief that p is true is irrelevant to whether she is justified in choosing φ.
3. Whether T's belief that p is 'Gettiered' is irrelevant to whether she is justified in choosing φ.
4. Knowledge is true, justified, UnGettiered belief.
5. So T is justfied in choosing φ iff she is justified in believing that p.
6. T is not justified in choosing φ.
7. So T is not justified in believing that p.

I think this argument is only plausible if we equivocate on what it is for a belief to be 'Gettiered'.

Assume first that 'Gettiered' means 'derived from a false intermediate step'. Then premise 4 is false, as Williamson's example shows. S has a justified true belief that is neither knowledge nor derived from a false premise.

Assume then that 'Gettiered' simply means that the true belief is justified without being known. In that case we have no reason to accept premise 3. After all, the class of true justified beliefs that are not knowledge is pretty open ended. Before reading Williamson, we may not have thought that this class included the beliefs of agents using measuring devices that were functioning properly but imperfectly. But it does. Prior to the end of epistemology, we simply don't know what other kind of beliefs might be in this class. There's no way to survey all the ways for justification to be insufficient for knowledge, and see if all of them are irrelevant to the justification for action. I think one way a justified belief can fall short of knowledge is if it is tied up with false beliefs about the stakes of bets. It's hard to say that that is irrelevant to the justification of action.

It is by now reasonably well known that logical subtraction is a very messy and complicated business. See, for instance, Humberstone (2000) for a clear discussion of the complications. In general, unless it is analytic that Fs are Gs and Hs, for some antecedently understood G and H, there's nothing interesting to say about the class of things that are G but not F. It will just be a disjunctive shambles. The same is true for knowledge and justification. The class of true beliefs that are justified but not known is messy and disjunctive. We shouldn't expect to have any neat way of overviewing it. That in part means we can't say much interesting about it as a class, contra premise 3 in the above argument. It also means the prospects for 'solving the Gettier problem' are weak. We'll turn to that issue next.

6.4 *There is No Solution to the Gettier Problem*

The kind of example that Edmund Gettier (1963) gives to refute the justified true belief theory of knowledge has what Linda Zagzebski (2009, 117) aptly calls a "double luck" structure. In Gettier's original cases, there's some bad luck that leads to a justified belief being false. But then there's some good luck that leads to an inference from that being true. As was quickly realised in the literature, the good and bad luck doesn't need to apply to separate inferential steps. It might be that the one belief that would have been false due to bad luck also ends up being true due to good luck.

This has led to a little industry, especially in the virtue epistemology section of the market, of attempts to "solve the Gettier problem" by adding an anti-luck condition to justification, truth and belief and hoping that the result is something like an analysis of knowledge. As Zagzebski (1994) showed, this can't be an *independent* condition on knowledge. If it doesn't entail truth, then we will be able to recreate the Gettier cases. But maybe a 'fourth' condition that entails truth (and perhaps belief) will suffice. Let's quickly review some of these proposals.

So Zagzebski (1996) suggested that the condition is that the belief be true *because* justified. John Greco (2010) says that the extra condition is that the beliefs be "intellectually creditable". That is, the primary that the subject ended up with a true belief is that it was the result of her reliable cognitive faculties. Ernest Sosa (2007) said that knowledge is belief that is true because it manifests intellectual competence. John Turri (2011) says that knowledge is belief the truth of which is a manifestation of the agent's intellectual competence.

It should be pretty clear that no such proposal can work if what I've said in earlier sections is remotely right. Assume again that $v = 830, a = 832$ and $m = 10$. The agent believes that $V \in [822, 842]$. This belief is, we've said, justified and true. Does it satisfy these extra conditions?

My short answer is that it does. My longer answer is that it does if any belief derived from the use of a measuring device does, and since some beliefs derived from the use of measuring devices amount to knowledge, the epistemologists are committed to the belief satisfying the extra condition. Let's go through those arguments in turn.

In our story, S demonstrates a range of intellectual competencies. She uses a well-functioning measuring device. It is the right kind of device for the purpose she is using. By hypothesis, she has had the machine carefully checked, and knows exactly the accuracy of the machine. She doesn't form any belief that is too precise to be justified by the machine. And she ends up with a true belief precisely because she has so many competencies.

Note that if we change the story so a is closer to $v+m$, the case that the belief is true in virtue of S being so competent becomes even stronger. Change the case so that $a = 839$, and she forms the true belief that $V \in [829, 849]$. Now if S had not been so competent, she may have formed a belief with a tighter range, since she could easily have guessed that the margin of error of the machine is smaller. So in this case the truth of the belief is very clearly due to her competence. But as we noted at the end of section 1, in the cases where a is near $v+m$, the argument that we have justified true belief without knowledge is particularly strong. Just when the gap between justification and knowledge gets most pronounced, the competence based approach to knowledge starts to issue the strongest verdicts *in favour* of knowledge.

But maybe this is all a mistake. After all, the object doesn't have the mass it has because of S's intellectual competence. The truth of any claim about its mass is not because of S's competence, or a manifestation of that competence. So maybe these epistemologists get the correct verdict that S does not know that $V \in [a-m, a+m]$?

Not so quick. Even had a equalled v, all these claims would have been true. And in that case, S would have known that V was within m of the measurement. What is needed for these epistemological theories to be right is that there can be a sense that a belief that p can be true in virtue of some cause C without C being a cause of p. I'm inclined to agree with the virtue epistemologists that such a sense can be given. (I think it helps to give up on content essentialism for this project, as suggested by David (2002) and endorsed inWeatherson (2004).) But I don't think it will help. There's no real way in which a belief is true because of competencies, or in which the truth of a belief manifests competence, in the good case where $a = v$, but not in the bad cases, where a is in $(0, m)$. These proposals from Zagzebski and others might help with explaining why a gap opens between knowledge and justification in 'double luck' cases. But that gap can appear in cases that don't have a 'double luck' structure. As noted in the previous section, I think the gap in question includes some cases involving false beliefs about the practical significance of p, but I don't expect everyone to agree with that. Happily, the Williamsonian cases should be less controversial.

Chapter 7

Objections and Replies

7.1 Interests are Defeaters

Interests can have a somewhat roundabout effect on knowledge. The IRI story goes something like this. If the agent has good but not completely compelling evidence for p, that is sometimes but not always sufficient for knowledge that p if everything else goes right. It isn't sufficient if they face a choice where the right thing to do is different to the right thing to do conditional on p. That is, if adding p to their background information would make a genuinely bad choice look like a good choice, they don't know that p. The reason is that if they did know p, they'd be in one of two unhappy states. The first such state is that they'd believe the choice is bad despite believing that conditional on something they believe, namely p, the choice is good. That's incoherent, and the incoherence defeats knowledge that p. The second such state is that they'd believe the choice is good. That's also irrational, and the irrationality defeats knowledge that p.[1]

Cases where knowledge is defeated because if the agent did know p, that would lead to problems elsewhere in their cognitive system, have a few quirky features. In particular, whether the agent knows p can depend on very distant features. Consider the following kind of case.

> **Confused Student**
>
> Con is systematically disposed to affirm the consequent. That is, if he notices that he believes both p and $q \rightarrow p$, he's disposed to either infer q, or if that's impermissible given his evidence, to ditch his belief in the conjunction of p and $q \rightarrow p$. Con has completely compelling evidence for both $q \rightarrow p$ and $\neg q$. He has good but less compelling evidence for p. And this evidence tracks the truth of p in

[1] Note again that I'm not here purporting to argue for IRI, just set out its features. Obviously not everyone will agree that knowledge that p can be defeated by irrationality or incoherence in closely related attitudes. But I think it is at least *plausible* that there are coherence constraints like this on knowledge, which is all I need for a defence of IRI against critics.

> just the right way for knowledge. On the basis of this evidence, Con believes p. Con has not noticed that he believes both p and $q \rightarrow p$. If he did, he's unhesitatingly drop his belief that p, since he'd realise the alternatives (given his dispositions) involved dropping belief in a compelling proposition. Two questions:
>
> - Does Con know that p?
> - If Con were to think about the logic of conditionals, and reason himself out of the disposition to affirm the consequent, would he know that p?

I think the answer to the first question is *No*, and the answer to the second question is *Yes*. As it stands, Con's disposition to affirm the consequent is a doxastic defeater of his putative knowledge that p. Put another way, p doesn't cohere well enough with the rest of Con's views for his belief that p to count as knowledge. To be sure, p coheres well enough with those beliefs by objective standards, but it doesn't cohere at all by Con's lights. Until he changes those lights, it doesn't cohere well enough to be knowledge.

I don't expect exactly everyone will agree with those judgments. Some people will simply reject that this kind of coherence by one's own lights is necessary for knowledge. Others might even reject the whole idea of doxastic defeaters. But I think the picture I've just sketched, one which puts reasonably tight coherence constraints on knowledge, is plausible enough to use in a defence of IRI. It certainly isn't so implausible that committing to it amounts to a *reductio* of one's views. Yet as we'll frequently see below, some criticisms of IRI do suggest that any theory that allows for these kinds of coherence constraints or doxastic defeaters is thereby shown to be false.[2] I'm going to take that suggestion to be a *reductio* of the criticisms.

7.2 IRI is an Existential Theory

IRI theorists do not typically say that interests are *always* relevant to knowledge. In fact, they hardly could be. If p is not true, or the agent has very little evidence for p, the agent does not know p whatever their interests. But an assumption that seems to shared by both some critics and some proponents of IRI is that IRI rests on some universal epistemic principles, not just on various existential principles. For instance, Jeremy Fantl and Matthew McGrath use a lot of principles like (JJ) in deriving IRI.

[2] This kind of criticism is most pronounced in the arguments I'll respond to in section 7.6, but it is somewhat pervasive.

(JJ) If you are justified in believing that p, then p is warranted enough to justify you in φ-ing, for any φ. (Fantl and McGrath, 2009, 99)

Now we it turns out, I think (JJ) is false. (I think it fails in cases where the agent is seriously mistaken about the risks and payoffs involved in doing φ, for instance.) But more importantly, we don't need anything nearly as strong as this to derive IRI. As long as there is *some* sufficiently large range of cases where (JJ) holds, we'll be able to establish the existence of *some* pair of cases which differ in whether the agent knows that p in virtue of the interests the agent has.

Relatedly, the argument for a version of IRI in Weatherson (2005a) makes frequent appeal to standard Bayesian decision theory. This might suggest that such an argument stands and falls with the success of consequentialism in decision theory. (I mean to use 'consequentialism' here roughly in the sense that it is used in Hammond (1988).) But again, this suggestion would be false. If consequentialism is true in some range of cases, we'll be able to use similar techniques to the ones used in that paper to show that there are the kinds of pairs of cases that IRI say exist.

7.3 Experimental Objections

As I mentioned in the discussion of the Map Cases, I don't think the argument for IRI rests on judgments, or intuitions, about similar cases. Rather, IRI can be independently motivated by, for instance, reflections on the relationship between belief and credence. It's a happy result, in my view, that IRI gets various Bank Cases and Map Cases right, but not essential to the view. If it turned out that the facts about the examples were less clear than we thought, that wouldn't *undermine* the argument for IRI, since those facts weren't part of the best arguments for IRI. But if it turned out that the facts about those examples were quite different to what IRI predicts, that may *rebut* the view, since it would then be shown to make false predictions.

This kind of rebuttal may be suggested by various recent experimental results, such as the results in May et al. (forthcoming) and Feltz and Zarpentine (2010). I'm going to concentrate on the latter set of results here, though I think that what I say will generalise to related experimental work.[3] In fact, I think the experiments don't really tell against IRI, because IRI doesn't make *any* unambiguous predictions about the cases at the centre of the experiments. The reason for this is related to the first point made in section one: it is odds, not stakes, that are most important.

[3] Note to editors: Because this work is not yet in press, I don't have page numbers for any of the quotes from Feltz and Zarpentine.

Feltz and Zarpentine gave subjects related vignettes, such as the following pair. (Each subject only received one of the pair.)

High Stakes Bridge John is driving a truck along a dirt road in a caravan of trucks. He comes across what looks like a rickety wooden bridge over a yawning thousand foot drop. He radios ahead to find out whether other trucks have made it safely over. He is told that all 15 trucks in the caravan made it over without a problem. John reasons that if they made it over, he will make it over as well. So, he thinks to himself, 'I know that my truck will make it across the bridge.'

Low Stakes Bridge John is driving a truck along a dirt road in a caravan of trucks. He comes across what looks like a rickety wooden bridge over a three foot ditch. He radios ahead to find out whether other trucks have made it safely over. He is told that all 15 trucks in the caravan made it over without a problem. John reasons that if they made it over, he will make it over as well. So, he thinks to himself, 'I know that my truck will make it across the bridge.' (Feltz and Zarpentine, 2010, ??)

Subjects were asked to evaluate John's thought. And the result was that 27% of the participants said that John does not know that the truck will make it across in **Low Stakes Bridge**, while 36% said he did not know this in **High Stakes Bridge**. Feltz and Zarpentine say that these results should be bad for interest-relativity views. But it is hard to see just why this is so.

Note that the change in the judgments between the cases goes in the direction that IRI seems to predict. The change isn't trivial, even if due to the smallish sample size it isn't statistically significant in this sample. But should a view like IRI have predicted a larger change? To figure this out, we need to ask three questions.

1. What are the costs of the bridge collapsing in the two cases?
2. What are the costs of not taking the bet, i.e., not driving across the bridge?
3. What is the rational credence to have in the bridge's sturdiness given the evidence John has?

IRI predicts that there is knowledge in Low Stakes Bridge but not in High Stakes Bridge only if the following equation is true:

$$\frac{C_H}{G + C_H} > x > \frac{C_L}{G + C_L}$$

where G is the gain the driver gets from taking a non-collapsing bridge rather than driving around (or whatever the alternative is), C_H is the cost of being on a collapsing bridge in High Stakes Bridge, C_L is the cost of being on a collapsing bridge in Low Stakes Bridge, and x is the probability that the bridge will collapse. I assume x is constant between the two cases. If that equation holds, then taking the bridge, i.e., acting as if the bridge is safe, maximises expected utility in Low Stakes Bridge but not High Stakes Bridge. So in High Stakes Bridge, adding the proposition that the bridge won't collapse to the agent's cognitive system produces incoherence, since the agent won't (at least rationally) act as if the bridge won't collapse. So if the equation holds, the agent's interests in avoiding C_H creates a doxastic defeater in High Stakes Bridge.

But does the equation hold? Or, more relevantly, did the subjects of the experiment believe that the equation hold? None of the four variables has their values clearly entailed by the story, so we have to guess a little as to what the subjects' views would be.

Feltz and Zarpentine say that the costs in "High Stakes Bridge [are] very costly—certain death—whereas the costs in Low Stakes Bridge are likely some minor injuries and embarrassment." (Feltz and Zarpentine, 2010, ??) I suspect both of those claims are wrong, or at least not universally believed. A lot more people survive bridge collapses than you may expect, even collapses from a great height.[4] And once the road below a truck collapses, all sorts of things can go wrong, even if the next bit of ground is only 3 feet away. (For instance, if the bridge collapses unevenly, the truck could roll, and the driver would probably suffer more than minor injuries.)

We aren't given any information as to the costs of not crossing the bridge. But given that 15 other trucks, with less evidence than John, have decided to cross the bridge, it seems plausible to think they are substantial. If there was an easy way to avoid the bridge, presumably the *first* truck would have taken it. If G is large enough, and C_H small enough, then the only way for this equation to hold will be for x to be low enough that we'd have independent reason to say that the driver doesn't know the bridge will hold.

But what is the value of x? John has a lot of information that the bridge will support his truck. If I've tested something for sturdiness two or three times, and it has worked, I won't even think about testing it again. Consider what evidence you need before you'll happily stand on a particular chair to reach something in the kitchen, or put a heavy television on a stand. Supporting a weight is the kind of thing that either fails the first time, or works fairly reliably. Obviously there

[4] In the West Gate bridge collapse in Melbourne in 1971, a large number of the victims were underneath the bridge; the people on top of the bridge had a non-trivial chance of survival. That bridge was 200 feet above the water, not 1000, but I'm not sure the extra height would matter greatly. Again from a slightly lower height, over 90% of people on the bridge survived the I-35W collapse in Minneapolis in 2007.

could be some strain-induced effects that cause a subsequent failure[5], but John really has a lot of evidence that the bridge will support him.

Given those three answers, it seems to me that it is a reasonable bet to cross the bridge. At the very least, it's no more of an unreasonable bet than the bet I make every day crossing a busy highway by foot. So I'm not surprised that 64% of the subjects agreed that John knew the bridge would hold him. At the very least, that result is perfectly consistent with IRI, if we make plausible assumptions about how the subjects would answer the three numbered questions above.

And as I've stressed, these experiments are only a problem for IRI if the subjects are reliable. I can think of two reasons why they might not be. First, subjects tend to massively discount the costs and likelihoods of traffic related injuries. In most of the country, the risk of death or serious injury through motor vehicle accident is much higher than the risk of death or serious injury through some kind of crime or other attack, yet most people do much less to prevent vehicles harming them than they do to prevent criminals or other attackers harming them.[6] Second, only 73% of this subjects in *this very experiment* said that John knows the bridge will support him in **Low Stakes Bridge**. This is just absurd. Unless the subjects endorse an implausible kind of scepticism, something has gone wrong with the experimental design. Given the fact that the experiment points broadly in the direction of IRI, and that with some plausible assumptions it is perfectly consistent with that theory, and that the subjects seem unreasonably sceptical to the point of unreliability about epistemology, I don't think this kind of experimental work threatens IRI.

7.4 Knowledge By Indifference and By Wealth

Gillian Russell and John Doris (2009) argue that Jason Stanley's account of knowledge leads to some implausible attributions of knowledge, and if successful their objections would generalise to other forms of IRI. I'm going to argue that Russell and Doris's objections turn on principles that are *prima facie* rather plausible, but which ultimately we can reject for independent reasons.[7]

Their objection relies on variants of the kind of case Stanley uses heavily in his (2005) to motivate a pragmatic constraint on knowledge. Stanley imagines a character who has evidence which would normally suffice for knowledge that p, but is faced with a decision where A is both the right thing to do if p is true, and

[5] As I believe was the case in the I-35W collapse.

[6] See the massive drop in the numbers of students walking or biking to school, reported in Ham et al. (2008), for a sense of how big an issue this is.

[7] I think the objections I make here are similar in spirit to those Stanley made in a comments thread on Certain Doubts, though the details are new. The thread is at http://el-prod.baylor.edu/certain_doubts/?p=616

will lead to a monumental material loss if *p* is false. Stanley intuits, and argues, that this is enough that they cease to know that *p*. I agree, at least as long as the gains from doing *A* are low enough that doing *A* amounts to a bet on *p* at insufficiently favourable odds to be reasonable in the agent's circumstance.

Russell and Doris imagine two kinds of variants on Stanley's case. In one variant the agent doesn't care about the material loss. As I'd put it, the agent's indifference to material odds shortens the odds of the bet. That's because costs and benefits of bets should be measured in something like utils, not something like dollars. Given that, Russell and Doris object that "you should have reservations ... about what makes [the knowledge claim] true: not giving a damn, however enviable in other respects, should not be knowledge-making." (Russell and Doris, 2009, 432). Their other variant involves an agent with so much money that the material loss is trifling to them. Again, this lowers the effective odds of the bet, so by my lights they may still know that *p*. But this is somewhat counter-intuitive. As Russell and Doris say, "[m]atters are now even dodgier for practical interest accounts, because *money* turns out to be knowledge making." (Russell and Doris, 2009, 433) And this isn't just because wealth can purchase knowledge. As they say, "money may buy the *instruments* of knowledge ... but here the connection between money and knowledge seems rather too direct." (Russell and Doris, 2009, 433)

The first thing to note about this case is that indifference and wealth aren't really producing knowledge. What they are doing is more like defeating a defeater. Remember that the agent in question had enough evidence, and enough confidence, that they would know *p* were it not for the practical circumstances. As I argued in section 7.1, practical considerations enter debates about knowledge in part because they are distinctive kinds of defeaters. It seems that's what is going on here. And we have, somewhat surprisingly, independent evidence to think that indifference and wealth do matter to defeaters.

Consider two variants on Gilbert Harman's 'dead dictator' example (Harman, 1973, 75). In the original example, an agent reads that the dictator has died through an actually reliable source. But there are many other news sources around, such that if the agent read them, she would lose her belief. Even if the agent doesn't read those sources, their presence can constitute defeaters to her putative knowledge that the dictator died.

In our first variant on Harman's example, the agent simply does not care about politics. It's true that there are many other news sources around that are ready to mislead her about the dictator's demise. But she has no interest in looking them up, nor is she at all likely to look them up. She mostly cares about sports, and will spend most of her day reading about baseball. In this case, the misleading news sources are too distant, in a sense, to be defeaters. So she still knows the dictator has died. Her indifference towards politics doesn't generate knowledge - the original reliable report is the knowledge generator - but her indifference means that a would-be defeater doesn't gain traction.

In the second variant, the agent cares deeply about politics, and has masses of wealth at hand to ensure that she knows a lot about it. Were she to read the misleading reports that the dictator has survived, then she would simply use some of the very expensive sources she has to get more reliable reports. Again this suffices for the misleading reports not to be defeaters. Even before the rich agent exercises her wealth, the fact that her wealth gives her access to reports that will correct for misleading reports means that the misleading reports are not actually defeaters. So with her wealth she knows things she wouldn't otherwise know, even before her money goes to work. Again, her money doesn't generate knowledge – the original reliable report is the knowledge generator – but her wealth means that a would-be defeater doesn't gain traction.

The same thing is true in Russell and Doris's examples. The agent has quite a bit of evidence that p. That's why she knows p. There's a potential practical defeater for p. But due to either indifference or wealth, the defeater is immunised. Surprisingly perhaps, indifference and/or wealth can be the difference between knowledge and ignorance. But that's not because they can be in any interesting sense 'knowledge makers', any more than I can make a bowl of soup by preventing someone from tossing it out. Rather, they can be things that block defeaters, both when the defeaters are the kind Stanley talks about, and when they are more familiar kinds of defeaters.

7.5 Temporal Embeddings

Michael Blome-Tillmann (2009) has argued that tense-shifted knowledge ascriptions can be used to show that his version of Lewisian contextualism is preferable to IRI. Like Russell and Doris, his argument uses a variant of Stanley's Bank Cases.[8] Let O be that the bank is open Saturday morning. If Hannah has a large debt, she is in a high-stakes situation with respect to O. In Blome-Tillman's version of the example, Hannah had in fact incurred a large debt, but on Friday morning the creditor waived this debt. Hannah had no way of anticipating this on Thursday. She has some evidence for O, but not enough for knowledge if she's in a high-stakes situation. Blome-Tillmann says that this means after Hannah discovers the debt waiver, she could say (2).

(2) I didn't know O on Thursday, but on Friday I did.

[8] In the interests of space, I won't repeat those cases yet again here.

But I'm not sure why this case should be problematic for any version of IRI, and very unsure why it should even look like a *reductio* of IRI. As Blome-Tillmann notes, it isn't really a situation where Hannah's stakes change. She was never actually in a high stakes situation. At most her perception of her stakes change; she thought she was in a high-stakes situation, then realised that she wasn't. Blome-Tillmann argues that even this change in perceived stakes can be enough to make (2) true if IRI is true. Now actually I agree that this change in perception could be enough to make (2) true, but when we work through the reason that's so, we'll see that it isn't because of anything distinctive, let alone controversial, about IRI.

If Hannah is rational, then given her interests she won't be ignoring $\neg O$ possibilities on Thursday. She'll be taking them into account in her plans. Someone who is anticipating $\neg O$ possibilities, and making plans for them, doesn't know O. That's not a distinctive claim of IRI. Any theory should say that if a person is worrying about $\neg O$ possibilities, and planning around them, they don't know O. And that's simply because knowledge requires a level of confidence that such a person simply does not show. If Hannah is rational, that will describe her on Thursday, but not on Friday. So (2) is true not because Hannah's practical situation changes between Thursday and Friday, but because her psychological state changes, and psychological states are relevant to knowledge.

What if Hannah is, on Thursday, irrationally ignoring $\neg O$ possibilities, and not planning for them even though her rational self wishes she were planning for them? In that case, it seems she still believes O. After all, she makes the same decisions as she would as if O were sure to be true. But it's worth remembering that if Hannah does irrationally ignore $\neg O$ possibilities, she is being irrational with respect to O. And it's very plausible that this irrationality defeats knowledge. That is, you can't be irrational with respect to a proposition and know it. Irrationality excludes knowledge. That's what we saw in Con's case in section 7.1, and it's all we see here as well. Note also that Con will know p after fixing his consequent-affirming disposition but not before. That's just what happens with Hannah; a distant change in her cognitive system will remove a defeater, and after it does she gets more knowledge.

There's a methodological point here worth stressing. Doing epistemology with imperfect agents often results in facing tough choices, where any way to describe a case feels a little counterintuitive. If we simply hew to intuitions, we risk being led astray by just focussing on the first way a puzzle case is described to us. But once we think through Hannah's case, we see perfectly good reasons, independent of IRI, to endorse IRI's prediction about the case.

7.6 Problematic Conjunctions

George and Ringo both have $6000 in their bank accounts. They both are thinking about buying a new computer, which would cost $2000. Both of them also have rent due tomorrow, and they won't get any more money before then. George lives in New York, so his rent is $5000. Ringo lives in Syracuse, so his rent is $1000. Clearly, (3) and (4) are true.

(3) Ringo has enough money to buy the computer.
(4) Ringo can afford the computer.

And (3) is true as well, though there's at least a reading of (4) where it is false.

(3) George has enough money to buy the computer.
(4) George can afford the computer.

Focus for now on (3). It is a bad idea for George to buy the computer; he won't be able to pay his rent. But he has enough money to do so; the computer costs $2000, and he has $6000 in the bank. So (3) is true. Admittedly there are things close to (3) that aren't true. He hasn't got enough money to buy the computer and pay his rent. You might say that he hasn't got enough money to buy the computer given his other financial obligations. But none of this undermines (3). The point of this little story is to respond to another argument Blome-Tillmann offers against IRI. Here is how he puts the argument. (Again I've changed the numbering and some terminology for consistency with this paper.)

> Suppose that John and Paul have exactly the same evidence, while John is in a low-stakes situation towards *p* and Paul in a high-stakes situation towards *p*. Bearing in mind that IRI is the view that whether one knows *p* depends on one's practical situation, IRI entails that one can truly assert:
>
> (2) John and Paul have exactly the same evidence for *p*, but only John has enough evidence to know *p*, Paul doesn't.
>
> (Blome-Tillmann, 2009, 328-9)

And this is meant to be a problem, because (2) is intuitively false.

But IRI doesn't entail any such thing. Paul does have enough evidence to know that *p*, just like George has enough money to buy the computer. Paul can't know that *p*, just like George can't buy the computer, because of their practical situations. But that doesn't mean he doesn't have enough evidence to know it. So, contra Blome-Tillmann, IRI doesn't entail this problematic conjunction.

In a footnote attached to this, Blome-Tillmann tries to reformulate the argument.

> I take it that having enough evidence to 'know *p*' in *C* just means having evidence such that one is in a position to 'know *p*' in *C*, rather than having evidence such that one 'knows *p*'. Thus, another way to formulate (2) would be as follows: 'John and Paul have exactly the same evidence for *p*, but only John is in a position to know *p*, Paul isn't.' (Blome-Tillmann, 2009, 329n23)

The 'reformulation' is obviously bad, since having enough evidence to know *p* isn't the same as being in a position to know it, any more than having enough money to buy the computer puts George in a position to buy it. But might there be a different problem for IRI here? Might it be that IRI entails (2), which is false?

(2) John and Paul have exactly the same evidence for *p*, but only John is in a position to know *p*, Paul isn't.

There isn't a problem with (2) because almost any epistemological theory will imply that conjunctions like that are true. In particular, any epistemological theory that allows for the existence of defeaters to not supervene on the possession of evidence will imply that conjunctions like (2) are true. Again, it matters a lot that IRI is suggesting that traditional epistemologists did not notice that there are distinctively pragmatic defeaters. Once we see that, we'll see that conjunctions like (2) are not surprising at all.

Consider again Con, and his friend Mod who is disposed to reason by modus ponens and not by affirming the consequent. We could say that Con and Mod have the same evidence for *p*, but only Mod is in a position to know *p*. There are only two ways to deny that conjunction. One is to interpret 'position to know' so broadly that Con is in a position to know *p* because he could change his inferential dispositions. But then we might as well say that Paul is in a position to know *p* because he could get into a different 'stakes' situation. Alternatively, we could say that Con's inferential dispositions count as a kind of evidence against *p*. But that stretches the notion of evidence beyond a breaking point. Note that we didn't say Con had any *reason* to affirm the consequent, just that he does. Someone might adopt, or change, a poor inferential habit because they get new evidence. But they need not do so, and we shouldn't count their inferential habits as evidence they have.

If that case is not convincing, we can make the same point with a simple Gettier-style case.

> **Getting the Job**
>
> In world 1, at a particular workplace, someone is about to be promoted. Agnetha knows that Benny is the management's favourite choice for the promotion. And she also knows that Benny is Swedish. So she comes to believe that the promotion will go to someone Swedish. Unsurprisingly, management does choose Benny, so Agnetha's belief is true.
>
> World 2 is similar, except there it is Anni-Frid who knows that Benny is the management's favourite choice for the promotion, that Benny is Swedish. So *she* comes to believe that the promotion will go to someone Swedish. But in this world Benny quits the workplace just before the promotion is announced, and the management unexpectedly passes over a lot of Danish workers to promote another Swede, namely Björn. So Anni-Frid's belief that the promotion will go to someone Swedish is true, but not in a way that she could have expected.

In that story, I think it is clear that Agnetha and Anni-Frid have exactly the same evidence that the job will go to someone Swedish, but only Agnetha is in a position to know this, Anni-Frid is not. The fact that an intermediate step is false in Anni-Frid's reasoning, but not Agnetha's, means that Anni-Frid's putative knowledge is defeated, but Agnetha's is not. And when that happens, we can have differences in knowledge without differences in evidence. So it isn't an argument against IRI that it allows differences in knowledge without differences in evidence.

7.7 Holism and Defeaters

The big lesson of the last few sections is that interests create defeaters. Sometimes an agent can't know p because adding p to her stock of beliefs would introduce either incoherence or irrationality. The reason is normally that the agent faces some decision where it is, say, bad to do φ, but good to do φ given p. In that situation, if she adds p, she'll either incoherently think that it's bad to do φ although it's good to do it given what is (by her lights) true, or she'll irrationally think that it's good to do φ. Moreover, the IRI theorist says, being either incoherent or irrational in this way blocks knowledge, so the agent doesn't know p.

But there are other, more roundabout, ways in which interests can mean that believing p would entail incoherence or irrationality. One of these is illustrated by an example alleged by Ram Neta to be hard for interest-relative theorists to accommodate.

> Kate needs to get to Main Street by noon: her life depends upon it. She is desperately searching for Main Street when she comes to an intersection and looks up at the perpendicular street signs at that intersection. One street sign says "State Street" and the perpendicular street sign says "Main Street." Now, it is a matter of complete indifference to Kate whether she is on State Street–nothing whatsoever depends upon it. (Neta, 2007, 182)

Let's assume for now that Kate is rational; dropping this assumption introduces mostly irrelevant complications.[9] Kate will not believe she's on Main Street. She would only have that belief if she took it to be settled that she's on Main, and hence not worthy of spending further effort investigating. But presumably she won't do that. The rational thing for her to do is to get confirming (or, if relevant, confounding) evidence for the appearance that she's on Main. If it were settled that she was on Main, the rational thing to do would be to try to relax, and be grateful that she had found Main Street. Since she has different attitudes about what to do *simpliciter* and conditional on being on Main Street, she doesn't believe she's on Main Street.

So far so good, but what about her attitude towards the proposition that she's on State Street? She has enough evidence for that proposition that her credence in it should be rather high. And no practical issues turn on whether she is on State. So she believes she is on State, right?

Not so fast! Believing that she's on State has more connections to her cognitive system than just producing actions. Note in particular that street signs are hardly basic epistemic sources. They are the kind of evidence we should be 'conservative' about in the sense of Pryor (2004). We should only use them if we antecedently believe they are accurate. So for Kate to believe she's on State, she'd have to believe street signs around here are accurate. If not, she'd incoherently be relying on a source she doesn't trust, even though it is not a basic source.[10] But if she believes the street signs are accurate, she'd believe she was on Main, and

[9] It means we constantly have to run through both the irrationality horn of the dilemma from the first paragraph of this section as well as the incoherence horn, but the two horns look very similar in practice.

[10] The caveats here about basic sources are to cancel any suggestion that Kate has to antecedently believe that any source is reliable before she uses it. As Pryor (2000) notes, that view is problematic. The view that we only get knowledge from a street sign if we antecedently have reason to trust it is not so implausible.

that would lead to practical incoherence. So there's no way to coherently add the belief that she's on State Street to her stock of beliefs. So she doesn't know, and can't know, that she's either on State or on Main. This is, in a roundabout way, due to the high stakes Kate faces.

Neta thinks that the best way for the interest-relative theorist to handle this case is to say that the high stakes associated with the proposition that Kate is on Main Street imply that certain methods of belief formation do not produce knowledge. And he argues, plausibly, that such a restriction will lead to implausibly sceptical results. But that's not the right way for the interest-relative theorist to go. What they should say is that Kate can't know she's on State Street because the only grounds for that belief is intimately connected to a proposition that, in virtue of her interests, she needs very large amounts of evidence to believe.

7.8 Non-Consequentialist Cases

None of the replies yet have leaned heavily on the point from section 7.2, the fact that IRI is an existential claim. This reply will make heavy use of that fact.

If an agent is merely trying to get the best outcome for themselves, then it makes sense to represent them as a utility maximiser. But when agents have to make decisions that might involve them causing harm to others if certain propositions turn out to be true, then I think it is not so clear that orthodox decision theory is the appropriate way to model the agents. That's relevant to cases like this one, which Jessica Brown has argued are problematic for the epistemological theories John Hawthorne and Jason Stanley have recently been defending.[11]

> A student is spending the day shadowing a surgeon. In the morning he observes her in clinic examining patient A who has a diseased left kidney. The decision is taken to remove it that afternoon. Later, the student observes the surgeon in theatre where patient A is lying anaesthetised on the operating table. The operation hasn't started as the surgeon is consulting the patient's notes. The student is puzzled and asks one of the nurses what's going on:
>
> **Student**: I don't understand. Why is she looking at the patient's records? She was in clinic with the patient this morning. Doesn't she even know which kidney it is?

[11] The target here is not directly the interest-relativity of their theories, but more general principles about the role of knowledge in action and assertion. But it's important to see how IRI handles the cases that Brown discusses, since these cases are among the strongest challenges that have been raised to IRI.

> **Nurse**: Of course, she knows which kidney it is. But, imagine what it would be like if she removed the wrong kidney. She shouldn't operate before checking the patient's records. (Brown, 2008, 1144-1145)

It is tempting, but I think mistaken, to represent the surgeon's choice as follows. Let **Left** mean the left kidney is diseased, and **Right** mean the right kidney is diseased.

	Left	**Right**
Remove left kidney	1	-1
Remove right kidney	-1	1
Check notes	$1-\varepsilon$	$1-\varepsilon$

Here ε is the trivial but non-zero cost of checking the chart. Given this table, we might reason that since the surgeon knows that she's in the left column, and removing the left kidney is the best option in that column, she should remove the left kidney rather than checking the notes.

But that reasoning assumes that the surgeon does not have any obligations over and above her duty to maximise expected utility. And that's very implausible, since consequentialism is a fairly implausible theory of medical ethics.[12]

It's not clear exactly what the obligation the surgeon has. Perhaps it is an obligation to not just know which kidney to remove, but to know this on the basis of evidence she has obtained while in the operating theatre. Or perhaps it is an obligation to make her belief about which kidney to remove as sensitive as possible to various possible scenarios. Before she checked the chart, this counterfactual was false: *Had she misremembered which kidney was to be removed, she would have a true belief about which kidney was to be removed.* Checking the chart makes that counterfactual true, and so makes her belief that the left kidney is to be removed a little more sensitive to counterfactual possibilities.

However we spell out the obligation, it is plausible given what the nurse says that the surgeon has some such obligation. And it is plausible that the 'cost' of violating this obligation, call it δ, is greater than the cost of checking the notes. So here is the decision table the surgeon faces.

[12] I'm not saying that consequentialism is wrong as a theory of medical ethics. But if it is right, so many intuitions about medical ethics are going to be mistaken that such intuitions have no evidential force. And Brown's argument relies on intuitions about this case having evidential value. So I think for her argument to work, we have to suppose non-consequentialism about medical ethics.

	Left	**Right**
Remove left kidney	$1-\delta$	$-1-\delta$
Remove right kidney	$-1-\delta$	$1-\delta$
Check notes	$1-\varepsilon$	$1-\varepsilon$

And it isn't surprising, or a problem for an interest-relative theory of knowledge, that the surgeon should check the notes, even if she believes *and knows* that the left kidney is the diseased one.

There is a very general point here. The best arguments for IRI start with the role that knowledge plays in a particular theory of decision or reasoning. It's easiest to make the arguments for IRI work if that theory is orthodox (consequentialist) decision theory. That doesn't mean that the arguments for IRI presuppose that consequentialism is always the right decision theory. As long as consequentialism is correct *in the case described*, the argument for IRI can work. (By consequentialism being correct in a case, I mean that it can be preferable to choose φ over ψ in some case because φ has the higher expected utility. It's plausible that there are such cases because it's plausible that there are choices we face where the options differ in no normatively salient respect except expected utility.) Remember, the IRI theorist is trying to prove an existential: there is a pair of cases that differ with respect to knowledge in virtue of differing with respect to interests. And that just needs consequentialism to be locally true. The only way medical cases like Brown's could be counterexamples to IRI is if we assumed that consequentialism was globally true, but it probably isn't, so IRI survives the examples.

Bibliography

Aumann, Robert J. 1999. "Interactive Epistemology I: Knowledge." *International Journal of Game Theory* 28:263–300.

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

Block, Ned. 1978. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9:261–325.

Blome-Tillmann, Michael. 2009. "Contextualism, Subject-Sensitive Invariantism, and the Interaction of 'Knowledge'-Ascriptions with Modal and Temporal Operators." *Philosophy and Phenomenological Research* 79:315–331, doi:10.1111/j.1933-1592.2009.00280.x.

Bovens, Luc and Hawthorne, James. 1999. "The Preface, the Lottery, and the Logic of Belief." *Mind* 108:241–264.

Braddon-Mitchell, David and Jackson, Frank. 2007. *The Philosophy of Mind and Cognition, second edition*. Malden, MA: Blackwell.

Brown, Jessica. 2008. "Knowledge and Practical Reason." *Philosophy Compass* 3:1135–1152, doi:10.1111/j.1747-9991.2008.00176.x.

—. forthcoming. "Impurism, Practical Reasoning and the Threshold Problem." *Noûs* , doi:10.1111/nous.12008.

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

Cohen, Stewart. 1984. "Justification and Truth." *Philosophical Studies* 46:279–295.

David, Marian. 2002. "Content Essentialism." *Acta Analytica* 17:103–114.

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

Feltz, Adam and Zarpentine, Chris. 2010. “Do You Know More When It Matters Less?” *Philosophical Psychology* 23:683–706, doi:10.1080/09515089.2010.514572.

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

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.

Greco, John. 2010. *Achieving Knowledge*. Cambridge: Cambridge University Press.

Ham, Sandra A., Martin, Sarah, and Kohl III, Harold W. 2008. “Changes in the percentage of students who walk or bike to school-United States, 1969 and 2001.” *Journal of Physical Activity and Health* 5:205–215.

Hammond, Peter J. 1988. “Consequentialist Foundations for Expected Utility.” *Theory and Decision* 25:25–78, doi:10.1007/BF00129168.

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.

Hawthorne, John and Stanley, Jason. 2008. “Knowledge and Action.” *Journal of Philosophy* 105:571–90.

Humberstone, Lloyd. 2000. “Parts and Partitions.” *Theoria* 66:41–82, doi:10.1111/j.1755-2567.2000.tb01144.x.

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

Ichikawa, Jonathan. 2009. "Explaining Away Intuitions." *Studia Philosophica Estonica* 22:94–116.

Ichikawa, Jonathan and Jarvis, Benjamin. 2009. "Thought-experiment intuitions and truth in fiction." *Philosophical Studies* 142:221–246, doi:10.1007/s11098-007-9184-y.

Jackson, Frank. 1991. "Decision Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101:461–82.

Kaplan, Mark. 1996. *Decision Theory as Philosophy*. Cambridge: Cambridge University Press.

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

Kennett, Jeanette and Smith, Michael. 1996a. "Frog and Toad Lose Control." *Analysis* 56:63–73, doi:10.1111/j.0003-2638.1996.00063.x.

—. 1996b. "Philosophy and Commonsense: The Case of Weakness of Will." In Michaelis Michael and John O'Leary-Hawthorne (eds.), *The Place of Philosophy in the Study of Mind*, 141–157. Norwell, MA: Kluwer, doi:10.1017/CBO9780511606977.005.

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

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

—. 1982. "Logic for Equivocators." *Noûs* 16:431–441. Reprinted in *Papers in Philosophical Logic*, pp. 97-110.

—. 1994. "Reduction of Mind." In Samuel Guttenplan (ed.), *A Companion to the Philosophy of Mind*, 412–431. Oxford: Blackwell, doi:10.1017/CBO9780511625343.019. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 291-324.

Lipsey, R. G. and Lancaster, Kelvin. 1956-1957. "The General Theory of Second Best." *Review of Economic Studies* 24:11–32, doi:10.2307/2296233.

Maitra, Ishani. 2010. "Assertion, Norms and Games."

Malmgren, Anna-Sara. 2011. "Rationalism and the Content of Intuitive Judgements." *Mind* 120:263–327, doi:10.1093/mind/fzr039.

May, Joshua, Sinnott-Armstrong, Walter, Hull, Jay G., and Zimmerman, Aaron. forthcoming. "Practical Interests, Relevant Alternatives, and Knowledge Attributions: an Empirical Study." *Review of Philosophy and Psychology* , doi:10.1007/s13164-009-0014-3.

Neta, Ram. 2007. "Anti-intellectualism and the Knowledge-Action Principle." *Philosophy and Phenomenological Research* 75:180–187, doi:10.1111/j.1933-1592.2007.00069.x.

Pryor, James. 2000. "The Skeptic and the Dogmatist." *Noûs* 34:517–549, doi:10.1111/0029-4624.00277.

—. 2004. "What's Wrong with Moore's Argument?" *Philosophical Issues* 14:349–378.

Ross, Jacob and Schroeder, Mark. fort. "Belief, Credence, and Pragmatic Encroachment." *Philosophy and Phenomenological Research* , doi:10.1111/j.1933-1592.2011.00552.xForthcoming.

Runyon, Damon. 1992. *Guys & Dolls: The stories of Damon Runyon*. New York: Penguin.

Russell, Gillian and Doris, John M. 2009. "Knowledge by Indifference." *Australasian Journal of Philosophy* 86:429–437, doi:10.1080/00048400802001996.

Ryle, Gilbert. 1949. *The Concept of Mind*. New York: Barnes and Noble.

Sainsbury, Mark. 1996. "Vagueness, Ignorance and Margin for Error." *British Journal for the Philosophy of Science* 46:589–601.

Sorensen, Roy A. 1988. *Blindspots*. Oxford: Clarendon Press.

Sosa, Ernest. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Oxford: Oxford University Press.

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.

—. 1998. "Belief revision in games: forward and backward induction." *Mathematical Social Sciences* 36:31 – 56. ISSN 0165-4896, doi:10.1016/S0165-4896(98)00007-9.

—. 1999. "Extensive and strategic forms: Games and models for games." *Research in Economics* 53:293 – 319. ISSN 1090-9443, doi:10.1006/reec.1999.0200.

Stalnaker, Robert C. 1996. "Knowledge, Belief and Counterfactual Reasoning in Games." *Economics and Philosophy* 12:133–163, doi:10.1017/S0266267100004132.

Stanley, Jason. 2005. *Knowledge and Practical Interests*. Oxford University Press.

Sturgeon, Scott. 2008. "Reason and the Grain of Belief." *Noûs* 42:139–165, doi:10.1111/j.1468-0068.2007.00676.x.

Turri, John. 2011. "Manifest Failure: The Gettier Problem Solved." *Philosophers' Imprint* 11:1–11.

Watson, Gary. 1977. "Skepticism about Weakness of Will." *Philosophical Review* 86:316–339, doi:10.2307/2183785.

Weatherson, Brian. 2003. "What Good Are Counterexamples?" *Philosophical Studies* 115:1–31, doi:10.1023/A:1024961917413.

—. 2004. "Luminous Margins." *Australasian Journal of Philosophy* 82:373 – 383.

—. 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.

—. 2006. "Questioning Contextualism." In Stephen Cade Hetherington (ed.), *Epistemology Futures*, 133–147. Oxford: Oxford University Press.

—. 2011. "Knowledge, Bets and Interests." In Jessican Brown and Mikkel Gerken (eds.), *forthcoming volume on knowledge ascriptions*. Oxford: Oxford University Press.

—. 2012a. "Games and the Reason-Knowledge Principle." *The Reasoner* 6:6–8.

—. 2012b. "Knowledge, Bets and Interests." In Jessican Brown and Mikkel Gerken (eds.), *Knowledge Ascriptions*, 75–103. Oxford: Oxford University Press.

White, Roger. 2005. "Epistemic permissiveness." *Philosophical Perspectives* 19:445–459.

Williamson, Timothy. 1994. *Vagueness*. Routledge.

—. 1998. "Conditionalizing on Knowledge." *British Journal for the Philosophy of Science* 49:89–121, doi:10.1093/bjps/49.1.89.

—. 2000. *Knowledge and its Limits*. Oxford University Press.

—. 2007. *The Philosophy of Philosophy*. Blackwell Pub. Ltd.

—. 2012. "Models of Improbable Knowing." Presented at the 2012 CSMN/Arché epistemology conference, Lofoten Islands, May 2012.

Zagzebski, Linda. 1994. "The Inescapability of Gettier Problems." *Philosophical Quarterly* 44:65–73.

—. 1996. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press.

—. 2009. *On Epistemology*. Belmont, CA.: Wadsworth.