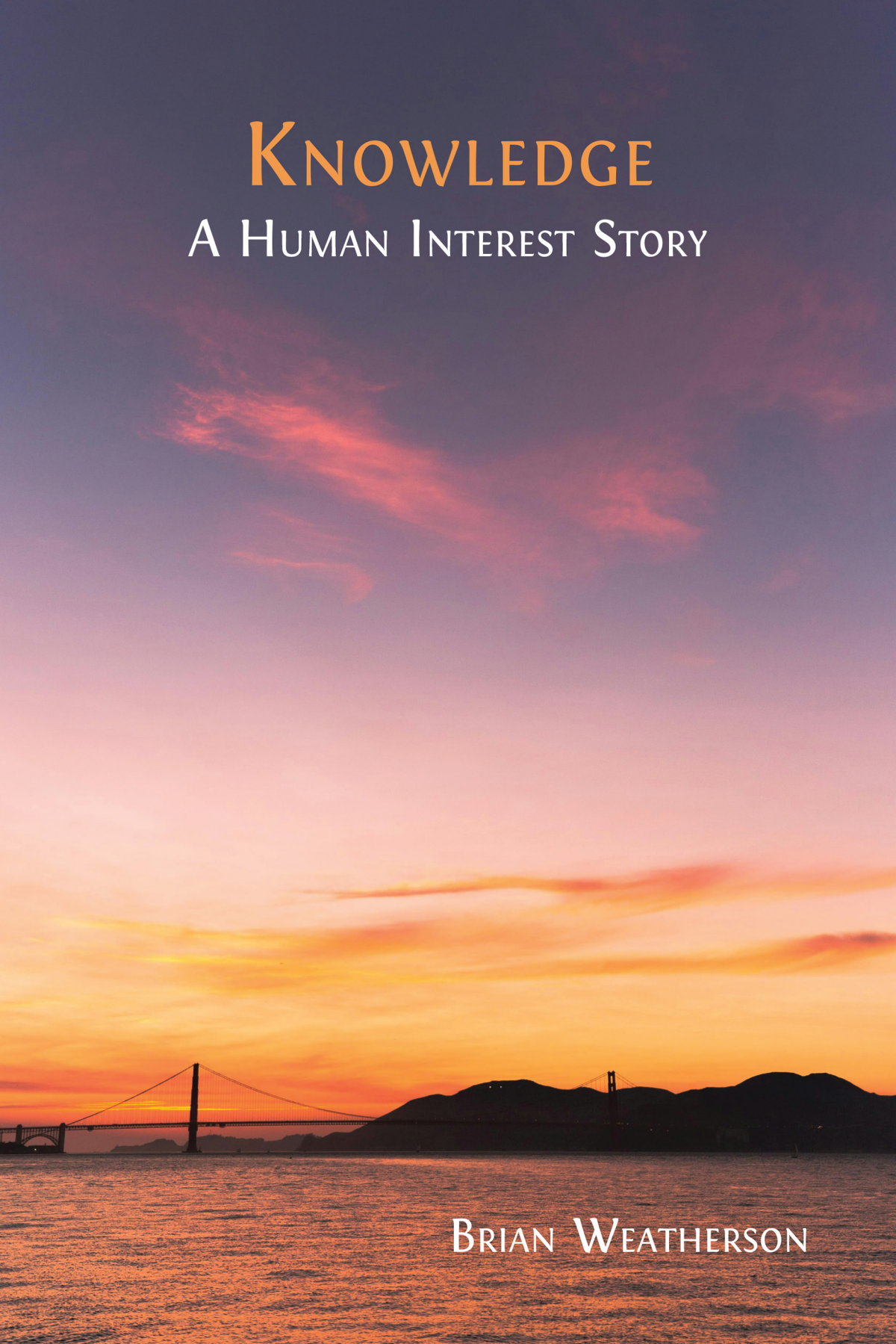
Knowledge

A Human Interest Story



Brian Weatherson

https://www.openbookpublishers.com

ISBN Paperback: 978-1-80511-394-2

ISBN Hardback: 978-1-80511-395-9

ISBN Digital (PDF): 978-1-80511-396-6

ISBN Digital eBook (EPUB): 978-1-80511-397-3

ISBN HTML: 978-1-80511-398-0

DOI: 10.11647/OBP.0425

Cover image: Golden Gate Bridge, San Francisco, California. Photo by Tj Kolesnik, https://unsplash.com/photos/silhouette-photography-of-mountain-beside-suspension-bridgeduring-golden-hour-Vnz7o7LeuDs.

Cover design: Jeevanjot Kaur Nagpal

9. Evidence

9.1 A Puzzle about Evidence

In Section 2.3.4, I argued that evidence can be interest-relative. The key example involved someone I called Parveen. Recall that she's in a restaurant and notices an old friend, Rahul, across the restaurant. The conditions for detecting people aren't perfect, and she's surprised Rahul is here. Still, we'd ordinarily say it is part of her evidence that Rahul is in this restaurant. She doesn't infer this from other facts, and she would not be called on to defend it if she relies on it in ordinary circumstances. She then plays the Red-Blue game, with these sentences.

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

The key premises for the argument that evidence is interest-relative are:

- The unique rational play for Parveen is Red-True.
- If evidence is interest-invariant, it is rational for Parveen to play Blue-True.

That argument shows that evidence is interest-relative. But it raises, without answering, two big questions:

1. When do interests matter for evidence?
2. When do interests matter for knowledge?

I used to think that there was an easy answer to the second question. A change in interest causes one to lose knowledge that p iff one becomes interested in a question which, given one's evidence, is rationally answered differently depending on whether or not one answers the question conditional on p. This answer is true as far as it goes, but it isn't

 https://doi.org/10.11647/OBP.0425.09

particularly explanatory unless one holds fixed the evidence between the earlier and later set of interests. And that is just what I said should not be held fixed.

The aim of this chapter is to answer both questions simultaneously.

9.2 A Simple, but Incomplete, Solution

To keep things relatively simple, I'll assume in this chapter that Parveen is an expected utility maximiser. More carefully, I'll assume that the reasons covered in Chapter 6 about why expected utility theory is only an approximation to the correct theory of rational choice are not relevant. From here on, we'll assume we're in a situation where expected utility theory is close enough to the true theory of rational choice.

At a very high level of abstraction, we can think about the problem facing Parveen (or anyone else whose evidence might be interest-sensitive), as follows. They have some option o, and given their interests it matters whether the expected value of o is above or below x. I'll write $v(\bullet)$ for the function from options to their expected value, so the question here is whether or not $v(o)$ us at least x.

There is some background K that is uncontroversially in Parveen's evidence. There is some further proposition p which might or might not be in her evidence; that's what the change of interests calls into question. It is uncontroversial that her evidence includes some background K, and controversial whether it includes some contested proposition p. For any q in K, $v(o \mid q) = v(o)$. That is, expected values are conditional on evidence.

A common idealisation helps capture this last idea. Assume there is a prior value function v^*, with a similar metaphysical status to the prior probability function. Then for any choice c, $v(c) = v^*(c \mid E)$, where E is the evidence Parveen has.

Now I can offer a simple, but incomplete, solution to question 2, assuming p is the only proposition whose status as evidence is put into question by the interests-shift, and the only shift in interests is that the question of whether $v(o) \geq x$ is now relevant. Then she knows p only if $[v^*(o|K) + v^*(o|K \wedge p)]/2 \geq x$. That is, if p's status as evidence is questionable, the relevant 'value' for o is the average of its expected value with and without p being evidence.

That gets the right answer about what Parveen should do. Her evidence may or may not include that Rahul is in the restaurant. If it does, then Blue-True has a value of $50. If it does not, then Blue-True's value is somewhat lower. Even if the evidence includes that someone who looks a lot like Rahul is in the restaurant, the value of Blue-True might only be $45. Averaging them out, the value is less than $50. It would only be rational to play Blue-True if was worth $50. So she shouldn't play Blue-True.

Great! Well, great except for two monumental problems. The first is that it only handles this very special case. The second is that the formula used, take the arithmetic mean of the values with and without the evidence, is barely better than arbitrary. It gets one thing right, in that it says Parveen shouldn't play Blue-True, but it's hardly alone in having that virtue.

Pragmatic encroachment starts with a very elegant, very intuitive, principle: you only know the things you can reasonably take to be settled for the purposes of current deliberation. This arbitrary averaging formula is not elegant or intuitive.

Happily, the two problems have a common solution. Setting it out requires going over recent work on coordination games.

9.3 The Radical Interpreter

William Harper (1986) pointed out that many decision problems are really better thought of as games. For instance, Newcomb's problem can be represented by the game in Table 9.1, with the human as Row and the demon as Column.

Table 9.1 Newcomb's problem as a game.

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

There is a unique equilibrium of this game: the bottom right corner. The reason it's the unique equilibrium is similar to the reason that two-boxers say to take two boxes: no other option is ratifiable for both players.

This section will be centred around a game that is only slightly more complicated. I call it The Interpretation Game. The game has two players. As in Newcomb's problem, they are a human and a mythical creature. Here the mythical creature is The Radical Interpreter.

In any game, the payouts are a function of what will happen to the players in each situation, and the players' values over those outcomes. To turn a physical situation into a game, we need to know the players' goals. Here are the goals I'll assume our players have:

- The Radical Interpreter assigns mental states to Human with the aim of making the action Human actually chooses the rational choice. I assume here that the 'mental states' include Human's evidence. Indeed, the main thing I'll have The Radical Interpreter do is assign evidence to Human.
- Human aims to maximise expected utility given their evidence. That last phrase, 'their evidence', should be read *de re*. More precisely, they aim to do the thing that is expected utility maximising given the evidence they actually have. (So their own views about their evidence don't matter; all that matters is what their evidence really is.)

Given these aims, The Radical Interpreter and Human often play coordination games. They will both achieve their aims if they act the 'same' way. That is, when it is uncertain whether p is part of Human's evidence, the coordination outcomes are:

- The Radical Interpreter says that p is part of Human's evidence, and Human maximises expected utility given $K \wedge p$.
- The Radical Interpreter says that p is part of Human's evidence, and Human maximises expected utility given K.

Coordination games typically have multiple equilibria, and that will also be the case here.

Let's focus on one example. Human is offered a bet on p. If the bet wins, it wins 1 util; if the bet loses, it loses 100 utils. Human's only choice is to Take or Decline the bet. The proposition p, the subject of the bet, is like the claim that Rahul is in the restaurant. That is, it is unclear whether it is in Human's evidence. Again, let K be the rest of Human's evidence, and stipulate that $\Pr(p \mid K) = 0.9$. Each party now faces a choice.

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

The payouts for the game are given in Table 9.2.

Table 9.2 The Radical Interpreter game.

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

Why is this the right table? Let's start with The Radical Interpreter.

The Radical Interpreter achieves their aim iff the following biconditional obtains: Human takes the bet iff p is part of their evidence. That's why they get payout 1 in the cells where that obtains, and 0 otherwise.

Most of Human's payouts are obvious. In the bottom row, they are guaranteed 0, since the bet is declined. In the top left, the bet wins with probability 1, so their expected return is 1. In the top right, the bet wins with probability 0.9, so the expected return of taking it is $1 \times 0.9 - 100 \times 0.1 = -9.1$.

There are two Nash equilibria for the game—the top left and the bottom right. We could stop here and say that according to IRT it is indeterminate whether p is part of Human's evidence. But we can do better.

But to do that, I need to survey more contested areas of game theory. In particular, I need to introduce some work on equilibrium choice. To do that, it helps to think about a game that is inspired by an example of Jean-Jacques Rousseau's.

9.4 Risk-Dominant Equilibria

Table 9.3 is the abstract version of a two-player, two-option game.

Table 9.3 A generic 2 by 2 by 2 game.

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

What are usually called Stag Hunt games have the following eight characteristics.

1. $r_{11} > r_{21}$
2. $r_{22} > r_{12}$
3. $c_{11} > c_{12}$
4. $c_{22} > c_{21}$
5. $r_{11} > r_{22}$
6. $c_{11} \geq c_{22}$
7. $r_{21} + r_{22} > r_{11} + r_{12}$
8. $c_{12} + c_{22} \geq c_{11} + c_{21}$

The first four conditions say that the game has two (strict) Nash equilibria: *Aa* and *Bb*. The next two conditions say that the *Aa* equilibrium is *Pareto-optimal*: neither player prefers *Aa* to *Bb*. In fact it says something a bit stronger: one of the players strictly prefers the *Aa* equilibrium, and the other player does not prefer *Bb*. The last two conditions say that the *Bb* equilibrium is *risk-optimal*.

Hans Carlsson and Eric van Damme (1993) offer an argument that in any such game, rational players will end up at *Bb*. The game that Human and The Radical Interpreter are playing fits these eight conditions, and The Radical Interpreter is perfectly rational. So if Carlsson and van Damme are right, The Radical Interpreter will say that $p \notin E$. Indeed, if Carlsson and van Damme are right, the toy theory I offered in Section 9.2 will be correct in all cases where it applies.

The rest of this chapter would be much simpler if I thought Carlsson and van Damme's argument worked in full generality. Unfortunately, I don't think it does. In particular, I think it fails in the important case where it is common knowledge that both players are rational, and both players know precisely the values of each of the eight payoffs. But I think it does work in the special case where one player has imperfect access to what the payouts are. And that, it turns out, is the special case that matters to us. That's getting ahead of the story though; let's start with their argument.

I said games satisfying these conditions are called Stag Hunt games. The name comes from a thought experiment in Rousseau's *Discourse on Inequality*.

> They were perfect strangers to foresight, and were so far from troubling themselves about the distant future, that they hardly thought of the morrow. If a deer was to be taken, every one saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs. (Rousseau, 1913: 209–210)

Brian Skyrms (2001) has argued that these Stag Hunt games are important across philosophy; they are good models for many real-life situations that are often (incorrectly) modelled as Prisoners' Dilemmas. But going over why that is would be a needless digression. Our focus is on Carlsson and van Damme's argument that Rousseau was right: a "stranger to foresight", who is just focussing on this game, should take the rabbit.

To make matters a little easier, we'll focus on a very particular instance of Stag Hunt, the one in Table 9.4.

Table 9.4 A simple version of Stag Hunt.

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

The equilibrium *Aa* is Pareto-optimal: it is the best outcome for each individual. But it is risky, and Carlsson and van Damme suggest a way to turn that risk into an argument for choosing *Bb*.

Embed Table 9.4 game in what they call a *global game*. Our first version of a global game is that each player knows that they will play Table 9.5, with x to be selected at random from a flat distribution over $[-1, 5]$.

Table 9.5 The global game.

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

There isn't much to say about Table 9.5 with this prior knowledge. Let's give the players a little more knowledge. (And we'll call the players Row and Column to make it easier to refer to each of them.)

Before they play the game, each player will get a noisy signal about the value of x. There will be signals s_R and s_C chosen (independently) from a flat distribution over $[x - 0.25, x + 0.25]$, and shown to Row and Column respectively. So each player will know the value of x to within ¼, and know that the other player knows it to within ¼ as well. This is a margin of error model, and in those models there is very little that is common knowledge. That, Carlsson and van Damme argue, makes a huge difference.

They go on to prove that iterated deletion of strictly dominated strategies (almost) removes all but one strategy pair. (I'll go over the proof of this in the next subsection.) Each player will play A/a if the signal is greater than 2, and B/b otherwise.[1] Surprisingly, this shows that players should play the risk-optimal strategy even when they know the other strategy is Pareto-optimal. When a player gets a signal in $(2, 3.75)$, then they know that $x < 4$, so *Bb* is the Pareto-optimal equilibrium. But the logic of the global game suggests the risk-dominant equilibrium is what to play.

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

Carlsson and van Damme go on to show that many of the details of this case don't matter. Most importantly, it doesn't matter that the margin of error in the signal was ¼; as long as it is positive the argument goes through.

Now what does this show about the game where players know precisely what the value of x is? Equivalently, what does it show about the game where the margin of error is 0?

Carlsson and van Damme argue that it shows that the risk-dominant choice is the right choice there as well. After all, the game where there is perfect knowledge just is a margin of error game, where the margin of error is 0. In previous work I'd endorsed this argument (Weatherson, 2018). I now think this was a mistake. The limit case, where the players know the value of x, is special. But, I'll argue, this doesn't actually undermine the argument that in the game between Human and The Radical Interpreter, both parties should choose the risk-dominant equilibria.

If the game between Human and The Radical Interpreter is meant to model a real situation, Human won't know precisely what the payoffs are. That's because real humans don't know precisely what their evidence is. They only know precisely what their evidence is if both positive and negative introspection hold for evidence, and that's no more plausible than that positive and negative introspection hold for knowledge. As Humberstone (2016: 380–402) shows, that's not particularly plausible, even if one doesn't accept the arguments in Williamson (2000) against positive introspection.

If Human doesn't know precisely what their evidence is, they don't know the payoffs in games like Table 9.5, because those payoffs are expected values. It turns out that's enough for the iterated dominance argument that Human should play the risk-dominant equilibrium to go through.

To be sure, The Radical Interpreter, who is just an idealisation, presumably does know the payouts in the different states of the game. It turns out, as I'll go over in Section 9.4.2, that Carlsson and van Damme's result only needs one player to be uncertain of the payouts. Given the failure of at least negative introspection (and, I'd say, positive introspection), that's something we can assume.

If Human should play the risk-dominant strategy in Table 9.2, they should decline the bet. So The Radical Interpreter, who can figure this out, should say that p is not part of their evidence. Since one's evidence just is what The Radical Interpreter says it is, that means that in Table 9.2, p is not part of Human's evidence.

Applied to the case of Parveen and Rahul, that means that The Radical Interpreter is best off saying it is no part of Parveen's evidence that Rahul is in the restaurant. More generally, in the simple cases described in Section 9.2, The Radical Interpreter should say that p is not part of Human's evidence just in case the equation used there holds.

The result is an interest-relative theory of evidence that is somewhat well motivated. At least, it can be incorporated into a broader theory of rational action.

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

The rest of this section goes into more technical detail about Carlsson and van Damme's example. Readers not interested in these details can skip ahead to Section 9.5. In Section 9.4.1 I summarise their argument that we only need iterated deletion of strictly dominated strategies to get the result that rational players will play the risk-dominant strategies. Then in Section 9.4.2 I offer a small generalisation of their argument, showing that it still goes through when one of the players gets a precise signal, and the other gets a noisy signal.

9.4.1 The Dominance Argument for Risk-Dominant Equilibria

Two players, Row (or R) and Column (or C) will play the game depicted in Table 9.5. They won't be told what x is, but they will get a noisy signal

of x, drawn from an even distribution over $[x - 0.25, x + 0.25]$. Call these signals s_R and s_C. Each player must then choose A, getting either 4 or 0 depending on the other player's choice, or choose B, getting x for sure.

Before getting the signal, the players must choose a strategy. In this context, a strategy is a function from signals to choices. Since the higher the signal is, the better it is to play B, we can more or less equate strategies with 'tipping points', where the player plays B if the signal is above the tipping point, and A below the tipping point.[2]

Call the tipping points for Row and Column respectively T_R and T_C. Since this game is symmetric, we'll just have to show that in conditions of common knowledge of rationality, $T_R = 2$. It follows by symmetry that $T_C = 2$ as well. The only rule that will be used is iterated deletion of strictly dominated strategies.

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

Note first that $T_R = 4.25$ strictly dominates any strategy where $T_R = y > 4.25$. If $s_R \in (4.25, y)$, then T_R is guaranteed to return above 4, and the alternative strategy is guaranteed to return 4. In all other cases, the strategies have the same return. There is some chance that $s_R \in (4.25, y)$. So we can delete all strategies $T_R = y > 4.25$, and similarly all strategies $T_C = y > 4.25$. By similar reasoning, we can rule out $T_R < -0.25$ and $T_C < -0.25$.

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

Now comes the important step. For arbitrary $y > 2$, assume we know that $T_C \leq y$. Consider the expected return of playing A given various values for $s_R > 2$. Given that the lower T_C is, the higher the expected return is of playing A, we'll just work on the simple case where $T_C = y$, realising that this is an upper bound on the expected return of A given $T_C \leq y$. The expected return of A is 4 times the probability that

2 I'm ignoring mixed strategies here, and strategies that differ in cases where the signal is right at the tipping point. It's trivial but tedious to extend the proof to cover these cases.

Column will play *a*, i.e., 4 times the probability that $s_C < T_C$. Given all the symmetries that have been built into the puzzle, we know that the probability that $s_C < s_R$ is 0.5. So the expected return of playing *A* is at most 2 if $s_R \geq y$. But the expected return of playing *B* is, as we showed in the last paragraph, *sR*, which is greater than 2. So it is better to play *B* than *A* if $s_R \geq y$. And the difference is substantial, so even if s_R is epsilon less than that *y*, it will still be better to play *B*. (This is rather hand-wavy, but I'll go over the more rigorous version presently.)

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

Let's make that a touch more rigorous. Assume that $T_C = y$, and s_R is slightly less than *y*. In particular, we'll assume that $z = y - s_R$ is in $(0, 0.5)$. Then the probability that $s_C < y$ is $0.5 + 2z - 2z^2$. So the expected return of playing *A* is $2 + 8z - 8z^2$. And the expected return of playing *B* is, again, *sR*. These will be equal iff $s_R = y + ((145 - 32y)^{½} - 9)/16$. So if we know that $T_C \geq y$, we know that $T_R \geq y + ((145 - 32y)^{½} - 9)/16$, which will be less than *y* if $y > 2$. Then by symmetry, we know that T_C must be at most as large as that as well. Then we can use that fact to derive a further upper bound on T_R and hence on T_C, and so on. And this will continue until we push both down to 2. It does require quite a number of steps of iterated deletion. Table 9.6 shows the upper bound on the threshold after *n* rounds of deletion of dominated strategies. (The numbers in Table 9.6 are precise for the first two rounds, and correct to three significant figures after that.)

Table 9.6 How the threshold moves towards 2.

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

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

Similar reasoning shows that at each stage not only are all strategies with higher tipping points dominated, but so are strategies that assign positive probability (whether it is 1 or less than 1), to playing A when the signal is above the 'tipping point'.[3]

So it has been shown that iterated deletion of dominated strategies will rule out all strategies except the risk-optimal equilibrium. The possibility that x is greater than the maximal return for A is needed to get the iterated dominance going. We also need the signal to have an

3 If we're careful about how we state this, we can use this to rule out all mixed strategies except those that respond probabilistically to $s_R = 2$.

error bar to it, so that each round of iteration removes more strategies. But that's all that was needed; the particular values used are irrelevant to the proof.

9.4.2 Making One Signal Precise

So far I've just been setting out Carlsson and van Damme's results. It's time to prove something just slightly stronger. I'll show that the result in Section 9.4.1 did not require that both parties receive a noisy signal. It's enough that just one party does.

More precisely, I'll change the game so that it is common knowledge that the signal Column gets, s_C, equals x. Since the game is no longer symmetric, I can't just appeal to the symmetry of the game as frequently as in the previous subsection. This slows the proof down, but doesn't stop it.

This change actually helps us at the first stage of the argument. Since Column could not be wrong about x, Column knows that if $s_C > 4$ then playing b dominates playing a. So one round of deleting dominated strategies rules out $T_C > 4$, as well as ruling out $T_R > 4.25$.

At any stage for any $y > 2$ such that we know $T_C \leq y$, the strategy $T_R = y$ dominates $T_R > y$. That's because if $s_R \geq y$, and $T_C \leq y$, the probability that Column will play a (given Row's signal) is less than 0.5. After all, the signal is just as likely to be above x as below it.[4] So if s_R is at or above T_C, the probability that Column's signal is above Column's tipping point is at least 0.5. So the probability that Column will play b is at least 0.5. So the expected return to Row of playing A, which is 4 times the probability that Column will play a, is at most 2. Since the expected return to Row of playing B equals the value of the signal,[5] that means that if the signal is above 2, they should play B.

Summing up, if Row knows $T_C \leq y$, for any $y > 2$, Row also knows it is better to play B if $s_R \geq y$. That is, if Row knows $T_C \leq y$, for any $y > 2$, Row's tipping point should be at most y.

4 This isn't strictly true if the signal is close enough to 5, but in that case we have an independent reason to think Column will play a.

5 Unless the signal is very close to 5, in which case they should play B anyway.

Assume now that it is common knowledge that $T_R \leq y$, for some $y > 2$. Assume Column's signal, which we'll call x, is just a little less than y. In particular, define $z = y - x$, and assume $z \in (0, 0.25)$. We want to work out the upper bound on the expected return to Column of playing a. (The return of playing b is known, it is x.)

The expected return to Column of playing a will be highest when T_R is highest. So we can work out an upper bound on that expected return by assuming that $T_R = y$. Given that assumption, the probability that Row plays A is $(1 + 2z)/2$. (That's the probability that Row's signal, which is a random draw from $[x - ¼, x + ¼]$, is above y.) So the expected return of playing a is $2 + 4z$, i.e., $2 + 4(y - x)$. That will be greater than x only when $x < (2 + 4y)/5$.

So if it is common knowledge that $T_R \leq y$, then it is best for Column to play b unless $x < (2 + 4y)/5$. That is, if it is common knowledge that $T_R \leq y$, then TC must be at most $(2 + 4y)/5$.

The rest of the proof proceeds in a zig-zag fashion. At one stage, we show that T_R must be no greater than T_C. So whatever value we've shown to be an upper bound for T_C is also an upper bound for T_R. At the next stage, we show that given any upper bound on T_R greater than 2, we can derive a new upper bound on T_C which is lower still. This process will eventually rule out all values for T_R and T_C greater than 2. So just using iterated deletion of dominated strategies, we eventually rule out all strategies that involve tipping points above 2.

There is one last point to be careful about. It takes infinitely many steps to rule out all tipping points above 2. Since it isn't obviously sound to have infinitely many steps of iterated deletion, one might worry about the soundness of the proof at this point. The key thing to note is that for any tipping point above 2, it is ruled out in a finite number of steps. So purely finitary reasoning rules out all tipping points above 2. It's just that there is no upper bound to the (finite!) number of steps needed.

This completes the mathematical part of the argument; I'll return to discussing whether this result matters for thinking about evidence and rational action, and reply to some objections to thinking that it does.

9.5 Objections and Replies

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

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

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

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

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

Reply: This is all true, and it shows that the view I'm putting forward is incompatible with some programs in epistemology. In particular, it is incompatible with E=K, since what it takes to be evidence in this story is slightly different from what it takes to be knowledge. The next section argues that this is independently plausible.

9.6 Evidence, Knowledge, and Cut-Elimination

In the previous section I noted that my theory of evidence is committed to denying Williamson's E=K thesis. This is the thesis that says one's evidence is all and only what one knows. What I say is consistent with, and arguably committed to, one half of that thesis. Nothing I've said here provides a reason to reject the implication that if p is part of one's evidence, then one knows p. Indeed, the story I'm telling would have to be complicated even further if that fails. But I am committed to denying the other direction. According to my view, there can be cases where someone knows p, but p is not part of their evidence.

My main reason for this comes from the kind of cases that Shyam Nair (2019) describes as failures of "cut-elimination". I'll quickly set out what Nair calls cut-elimination, and why it fails, and then look at how it raises problems for E=K.

Start by assuming that we have an operator $\models$ such that $\Gamma \models A$ means that A can be rationally inferred from Γ. I'm following Nair (and many others) in using a symbol usually associated with logical entailment here, though this is potentially misleading. A big plotline in what follows will be that $\models$, so understood, behaves very differently from familiar notions of entailment.

For the purposes of this section, I'm staying somewhat neutral on what it means to be able to rationally infer A from Γ. In particular, I want everything that follows to be consistent with the interpretation that an inference is rational only if it produces knowledge. I don't think that's true; I think folks with misleading evidence can rationally form false beliefs, and I think the traveller in Dharmottara's example rationally believes there is a fire. But there is a dialectical reason for staying neutral here. I'm arguing against one important part of the 'knowledge first' program, and I don't want to do so by assuming the falsity of other parts of it. So for this section (only), I'll write in a way that is consistent with saying rational belief requires knowledge.

Given that, one way to interpret $\Gamma \models A$ is that A can be known on the basis of Γ. What can be known on the basis of what is a function of, among other things, who is doing the knowing, what their background evidence is, what their capacities are, and so on. Strictly speaking, that suggests we should have some subscripts on $\models$ for who is the knower,

what their background evidence is, and so on. In the interests of readability, I'm going to leave all those implicit. In the next section it will be important to come back and look at whether the force of some of these arguments is diminished if we are careful about this relativisation.

That's our important notation. The principle *Cut* that Nair focuses on is that if 1 and 2 are true, so is 3.

9. $\Gamma \models A$
10. $\{A\} \cup \Delta \models B$
11. $\Gamma \cup \Delta \models B$

The principle is intuitive. Indeed, it is often implicit in a lot of reasoning. Here is one instance of it in action.

> I heard from a friend that Jack went up the hill. This friend is trustworthy, so I'm happy to infer that Jack did indeed go up the hill. I heard from another friend that Jack and Jill did the same thing. This friend is also trustworthy, so I'm happy to infer that Jill did the same thing as Jack, i.e., go up the hill.

Normally we wouldn't spell out the 'happy to infer' steps, but I've included them in here to make the reasoning a bit more explicit. But note what I didn't need to make explicit, even in this laborious reconstruction. I didn't need to note a change of status of the claim that Jack went up the hill. That goes from being a conclusion to being a premise. What matters for our purposes is that there doesn't seem to be a gap between the rationality of inferring that Jack went up the hill, and the rationality of using that as a premise in later reasoning. The idea that there is no gap here just is the idea that the principle *Cut* is true.

While *Cut* seems intuitive in cases like this, Nair argues that it can't be right in general. (If that's right we have a duty, one Nair takes up, to explain why cases like Jack and Jill seem like cases of good reasoning.) For my purposes, it is helpful to divide the putative counterexamples to *Cut* into two categories. I'll call them *monotonic* and *non-monotonic* counterexamples. The categorisation turns on whether $\Gamma \cup \Delta \models A$ is true assuming that $\Gamma \models A$ is true. I'll call cases where it is true monotonic instances of *Cut*, and cases where it is false non-monotonic instances.

That *Cut* fails in non-monotonic cases is fairly obvious. We can see this with an example that was hackneyed a generation ago.

Γ = {Tweety is a bird}
Δ = {Tweety is a penguin}
A = B = Tweety can fly

From Tweety is a bird we can rationally infer that Tweety flies. And given that Tweety is a flying penguin, we can infer that she flies. But given that Tweety is a penguin and a bird, we cannot infer this. So principles 1 and 2 in *Cut* are true, but 3 is false. And the same pattern will recur any time Δ provides a defeater for the link between Γ and *A*.

These cases will matter in what follows, but they are rather different from the monotonic examples. The monotonic example I'll set out (in the next three paragraphs) is very similar to one used in an argument against E=K by Alvin Goldman (2009). In many ways the argument against E=K I'm going to give is just a notational variant on Goldman's, but I think the notation I'm borrowing from Nair helps bring out the argument's strength.

Here is the crucial background assumption for the example. (I'll come back to how plausible this is after setting the example up.) The nature of *F* around here varies, but it varies very slowly. If we find a pattern in common to all the *F* within distance (in miles) *d* of here, we can rationally infer that the pattern extends another mile. That's just boring induction. But we can't infer that it extends to infinity, that would be a radical step. If we can't infer that the pattern goes to infinity, there must be a point beyond which we can't infer the pattern goes. Let's say that's one mile. So if we know the pattern holds within distance *d* of here, we can infer that it holds within distance $d + 1$, but no more.[6]

To see a case like this, imagine we're doing work that's more like working out the diet of local wildlife than working out the mass of an electron. If you know the mass of electrons around here, and what pigeons around here eat, there are some inferences you can make. You can come to know what the mass of electrons will be in the next town over, and what pigeons eat in the next town over. But there is a difference between the cases. You can also infer from this evidence what the mass of electrons will be on the other side of the world. But you can't make very confident inferences about what pigeons eat on the other side of

6 In any remotely realistic case, it would make more sense to say we can infer it holds in some multiple of *d* rather than adding some value to *d*. But I'm simplifying a lot to make a point, and this is just one more simplification.

the world; they may have adapted their diet to local conditions. In our case F and G concern things more like pigeon diets than electron masses.

Now here is the counterexample.

$\Gamma = \Delta =$ {Every F within 3 miles of here is G.}
$A =$ Every F between 3 and 4 miles of here is G.
$B =$ Every F between 4 and 5 miles of here is G.

If what I said was right, then this is a counterexample to *Cut*. $\Gamma \models A$ is true because it says given evidence about all the F within 3 miles of here, we can infer that all the F within 4 miles are like them. And $\{A\} \cup \Delta \models B$ is true because it says that given evidence about all the F within 4 miles of here, we can infer that all the F within 5 miles are like them. But $\Gamma \cup \Delta \models A$ is false, because it purports to say that given evidence about the F within 3 miles of here, we can infer that all the F within 5 miles are alike. And that's an inductive bridge too far.

This particular example involving distances was an extreme idealisation. But all we need for the larger argument is that there is some similarity metric such that inductive inference is rational across short jumps in that similarity metric, but not across long jumps. One kind of similarity is physical distance from a salient point. That's not the only kind of similarity, and rarely the most important kind.

As long as there is some 'inductive margin of inference', the argument works. What I mean by an inductive margin of inference is that given that all the F that differ from a salient point (along this metric) by amount d are G, it is rational to infer that all the F that differ from that salient point by amount $d + m$ are G, but not that all the F that differ from that salient point by amount $d + 2m$ are G. And it seems very plausible to me that there are some metrics, and values of F, G, d, m such that that's true.

For example, given what I know about Miami's weather, I can infer that it won't snow there for the next few hundred Christmases. Indeed, I know that. But I can't know that it won't snow there for the next few million Christmases. There is some point, and I don't know what it is, where my inductive knowledge about Miami's snowfall (or lack thereof) gives out.

While it is plausible that such cases are possible, any particular case fitting this pattern is weird. Here's what is weird about them. It will be easier to go back to the case where the metric is physical distance

to set this out, but the weirdness will extend to all cases. Imagine we investigate the area within 3 miles of here thoroughly, and find that all the *F* are *G*s. We infer, and now know, that all the *F* within 4 miles of here are *G*s. We keep investigating, and keep observing, and after a while we've observed all the *F* within 4 miles. And they are all *G*, as we knew they would be. But now we are in a position to infer that all the *F* within 5 miles are *G*. Observing something that we knew to be true gives us a reason to do something, i.e., make a further inference, that we couldn't do before. That's weird, and I'm going to come back in the next section to how it relates to the story I told about knowledge in Chapter 4.

The key point now is that this possibility undermines E=K. There is a difference between knowing *A* and being able to use *A* to support further inductive inferences. It is very natural to call that the difference between knowing *A* and having *A* as evidence.

The reasoning that I've been criticising violates a principle Jonathan Weisberg calls "No Feedback" (Weisberg, 2010: 533–534). This principle says that if a conclusion is derived from some premises, plus some intermediary conclusions, then it is only justified if it could, at least in principle, be derived from those premises alone. A natural way to read this is that we have some evidence, and things that we know on the basis of that evidence have a different functional role from the evidence. They can't do what the evidence itself can do, even if known. This looks like a problem for E=K, as Weisberg himself notes (2010: 536).

If any monotonic instances of failures of Cut exist, we need to distinguish between things the thinker knows by inference, and things they know by observation, in order to assess their inferences. That's to say, some knowledge will not play the characteristic role of evidence. T suggests that E=K is false.

9.7 Basic Knowledge and Non-Inferential Knowledge

It would be natural to conclude from the examples I've discussed that evidence is something like non-inferential knowledge. This is very similar to a view defended by Patrick Maher (1996). And it is, I will argue, close to the right view. But it can't be exactly right, for reasons Alexander Bird (2004) brings out.

I will argue that evidence is not non-inferential knowledge, but rather basic knowledge. The primary difference between these two notions is that *being non-inferential* is a diachronic notion—it depends on the causal source of the knowledge—while being basic is a synchronic notion—it depends on how the knowledge is currently supported. In general, non-inferential knowledge will be basic knowledge, and basic knowledge will be non-inferential. But the two notions can come apart, and when they do, the evidence is what is basic, not what is non-inferential.

The following kind of case is central to Bird's objection to the idea that evidence is non-inferential knowledge. Assume that our inquirer sees that A and rationally infers B. On the view that evidence is non-inferential knowledge, A is evidence but B is not. Now imagine that at some much later time, the inquirer remembers B, but has forgotten that it is based on A. This isn't necessarily irrational. As Gilbert Harman (1986) stresses, an obligation to remember our evidence is wildly unrealistic. The inquirer learns C and infers $B \wedge C$. This seems perfectly rational. But why is it rational?

If evidence is non-inferential knowledge, then this is a mystery. Since B was inferred, that can't be the evidence that justifies $B \wedge C$. So the only other option is that the evidence is the, now forgotten, A. It is puzzling how something that is forgotten can now justify. But a bigger problem is that if A is the inquirer's evidence, then they should also be able to infer $A \wedge C$. But this would be an irrational inference.

So I agree with Bird that we can't identify evidence with non-inferential knowledge, if by that we mean knowledge that was not originally gained through inference. (And what else could it mean?) But a very similar theory of evidence can work. The thing about evidence is that it can play a distinctive role in reasoning—it provides a distinctive kind of reason. In particular, it provides basic reasons.

Evidence stops regresses. That's why we can say that our fundamental starting points are self-evident. Now there is obviously a controversy about what things are self-evident. I don't find it particularly likely that claims about the moral rights we were endowed with by our Creator are self-evident. But I do think it is true that a lot of things are self-evident. (Even including, perhaps, that we have moral rights.) We should take this notion of self-evidence seriously. Sometimes a piece of knowledge

is a basic reason; it is evidence for itself, and not something that is grounded in further evidence.

What is it for a reason to be basic? It isn't that it was not originally inferred. Something that was once inferred from long forgotten premises may now be a basic reason. Rather, it is something that needs no further reason given as support. (Its support is itself, since it is self-evident.) What makes a reason need further support? I'm an interest-relative epistemologist, so I think this will be sensitive to the agent's interests. For example, I think facts reported in a reliable history book are pieces of basic evidence when we are thinking about history, but not when we are thinking about the reliability of that book. But this kind of interest-relativity is inessential to the story. What is essential is that evidence provides a reason that does not in turn require more justification.

This picture suggests an odd result about cases of forgotten evidence. There is a much-discussed puzzle about forgotten evidence that was set in motion by Harman (1986). He argued that if someone irrationally believes p on the basis of some evidence, then later forgets the evidence but retains the belief, the belief may now be rational. It would not be rational if they remembered both the evidence, and that it was the evidence for p. But, and this is what I want to take away from the case, there is no obligation for thinkers to keep track of why they believe each of the things they do.

There is a large literature now on this case; Sinan Dogramaci (2015) both provides a useful guide to the debate and moves it forward by considering what we might aim to achieve by offering one or other evaluation of the believer in this case. The view I'm offering here is, as far as I can tell, completely neutral on Harman's original case. But it has something striking to say about a similar case.

Imagine an inquirer, call him Jaidyn, believes p for the excellent reason that he read it in a book from a reliable historian H. Six months later, he has forgotten that that's where he learned that p, though he still believes that p. In a discussion about historians, a friend of Jaidyn's says that H is really unreliable. Jaidyn is a bit shocked, and literally can't believe it. This is for the best since H is in fact reliable, and his friend is suffering from a case of mistaken identity. But he is moved enough by the testimony to suspend judgment on H's reliability, and so he forms a disposition to not believe anything H says without corroboration. Since he doesn't know

that he believes *p* because *H* says so, he doesn't do anything about this belief. What should we say about Jaidyn's belief that *p*?

Here's what I want to say. I don't claim this is particularly intuitive, but I'm not sure there is anything particularly intuitive; it's best to just see what a theory says about the case. My theory says that Jaidyn still knows that *p*. This knowledge was once based on *H*'s testimony, but it is no longer based on that. Indeed, it is no longer based on anything. Presumably, if Jaidyn is rational, the knowledge will be sensitive to the absence of counter-evidence, or to incoherence with the rest of his world-view. But these are checks and balances in Jaidyn's doxastic system, they aren't the basis of the belief. Since the belief is knowledge, and is a basic reason for Jaidyn, it is part of his evidence.

Note three things about that last conclusion. First, this is a case where a piece of inferential knowledge can be in someone's evidence. By (reasonably) forgetting the source of the knowledge, it converts to being evidence. Second, almost any knowledge could make this jump. Whenever someone has no obligation to remember the source or basis of some knowledge, they can reasonably forget the source, and the basis, and the knowledge will become basic. And then it is evidence. The picture I'm working with is that pieces of knowledge can easily move in and out of one's evidence set; sometimes all it takes is forgetting where the knowledge came from. But third, if Jaidyn had done better epistemically, and remembered the source, he would no longer know that *p*.

It is somewhat surprising that knowledge can be dependent on forgetting. Jaidyn knows that *p*, but if he'd done better at remembering why he believes *p*, he wouldn't know it. Still, the knowledge isn't grounded in forgetting. It's originally grounded in testimony from an actually reliable source, and Jaidyn did as good a job as he needed to in checking the reliability of the source before accepting the testimony. Now since Jaidyn is finite, he doesn't have any obligation to remember everything. It seems odd to demand that Jaidyn adjust his beliefs on the basis of where they are from if he isn't even required to track where they are from. It would be very odd to say that Jaidyn's evidence now includes neither *p* (because it is undermined by his friend's testimony), nor the fact that someone said that *p*. That suggests any *p*-related

inferences Jaidyn makes are totally unsupported by his evidence, which doesn't seem right.

So the picture of evidence as basic knowledge, combined with a plausible theory of when forgetting is permissible, suggests that the forgetful reader knows more than the reader with a better memory. I suspect the same thing will happen in versions of Goldman's explosive inductive argument. Imagine a thinker observes all the *F*s within 3 miles, sees they are all *G*, and rationally infers that all the *F*s within 4 miles are *G*. Some time later they retain the belief, the knowledge actually, that all *F*s within 4 miles are *G*. But they forget that this was partially inferential knowledge, like Jaidyn forgot the source of his knowledge that *p*. They then make the seemingly sensible inductive inference that all *F*s within 5 miles are *G*. Is this rational, and can it produce knowledge? I think the answer is yes; if they (not unreasonably) forget the source of their knowledge that the *F*s 3 to 4 miles away are *G*, then this knowledge becomes basic. If it's basic, it is evidence. And if it is evidence, it can support one round of inductive reasoning.

I've drifted a fair way from discussing interest-relativity. And a lot of what I say here is inessential to defending IRT. So I'll return to the main plotline with a discussion of how my view of evidence helps respond to a challenge Ram Neta issues to IRT, and implies a rejection of a key principle in Jeremy Fantl and Matthew McGrath's theory of knowledge.

9.8 Holism and Defeaters

The picture of evidence I've outlined here grounds a natural response to a nice puzzle case outlined by Ram Neta (2007).[7]

> 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)

7 This section draws Weatherson (2011: §5).

Neta argues that IRT implies Kate knows that she is on State Street, but does not know that she is on Main Street. He suggests this is intuitively implausible. I think I agree with that intuition, so let's take it for granted and ask whether IRT has this problematic implication.

Let's also assume that it is not rational for Kate to take the street sign's word for it. I'm not sure that's true actually, but let's assume it to get the argument going. I think Neta is reasoning that since Kate's life depends on it, then IRT must say that she can't trust street signs, because the stakes are so high.

That claim about the relation between stakes and what one can take for granted can't be right. I often take actions that my life depends on going by the say so of signs. For example, I often turn onto the freeway ramp labelled 'on ramp', and not the ramp labelled 'off ramp', without really double checking. If I was wrong about this there is a very high chance I'd be very quickly killed. (Wrong-way crashes on freeways are a very common kind of fatal collisions.) If Kate can't take the sign for granted, it isn't just because her life is at stake; somewhat disconcertingly, that doesn't make the case any different from everyday driving.

But maybe Kate has some other way of checking where she is—like a map on a phone in her pocket—and it would be irrational to take the sign for granted and not check that other map. So I'm not going to push on this assumption.

So what evidence should The Radical Interpreter assign to Kate? It doesn't seem to be at issue that Kate sees that the signs say State and Main. The big question is whether she can simply take it as evidence that she is on State and Main. That is, do the contents of the sign simply become part of Kate's evidence? (Assume that the signs are accurate and there is no funny business going on, so it is plausible that the signs contribute to this evidence.) There are three natural options.

1. Both signs supply evidence directly to Kate, so her evidence includes that she is on State and that she is on Main.
2. Neither sign contributes evidence directly to Kate, so her evidence includes what the signs say, but nothing directly about her location.
3. One sign contributes evidence directly to Kate, but the other does not.

Option 1 implies that Kate is rational to not check further whether she is on Main Street. And that's irrational, so option 1 is out.

Option 3 implies that the signs behave differently, and that The Rational Interpreter will assign them different roles in Kate's cognitive architecture. But this will be true even though the signs are equally reliable, and Kate's evidence about their reliability is identical. So Kate treating them differently would be irrational, and The Radical Interpreter does not want to make Kate irrational if it can be helped. So option 3 is out.

That leaves option 2. Kate's evidence does not include that she is on State, and does not include that she is on Main. The latter 'non-inclusion' is directly explained by pragmatic factors. The former is explained by those factors plus the requirement that Kate's evidence is what The Radical Interpreter says it is, and The Radical Interpreter's desire to make Kate rational.

So Kate's evidence doesn't distinguish between the streets. It does, however, include that the signs say she is on State and that she is on Main. Could she be justified in inferring that she is on State, but not that she is on Main?

It is hard to see how this could be so. Street signs are hardly basic epistemic sources. They are the kind of evidence we should be 'conservative' about in the sense of James Pryor (2004). We should only use them if we antecedently believe they are correct. So for Kate to believe she's on State, she'd have to believe the street signs she can see are correct. If not, she'd incoherently be relying on a source she doesn't trust, even though it is not a basic source. But if she believes the street signs are correct, she'd believe she was on Main, and that would lead to practical irrationality. 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 practical situation Kate faces.

Neta thinks that the best way for IRT 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.

What to say about this suggestion turns on how we understand what a 'method' is. If methods are individuated very finely, like *Trust street signs right here*, then it's plausible that Kate should restrict what methods she uses, but implausible that this is badly sceptical. If methods are individuated very coarsely, like *Trust written testimony*, then it's plausible that this is badly sceptical, but implausible that Kate should give up on methods this general. I can rationally treat some parts of a book as providing direct evidence about the world, and other, more speculative, parts as providing direct evidence about what the author says, and hence indirect evidence about the world. Similarly, Kate can treat these street signs as indirect evidence about her location, while still treating other signs around her as providing direct evidence. So there is no sceptical threat here.

But while the case doesn't show IRT is false, it does tell us something interesting about the implications of IRT. When a practical consideration defeats a claim to know that p, it will often also knock out nearby knowledge claims. Some of these are obvious, like that the practical consideration defeats the claim to know $0{=}0 \rightarrow p$. But some of these are more indirect. When the inquirer knows what her evidence is, and knows that she has just the same evidence for q as for p, then if a practical consideration defeats a claim to know p, it also defeats a claim to know q. In practice, this makes IRT a somewhat more sceptical theory than it may have first appeared. It's not so sceptical as to be implausible, but it's more sceptical than is immediately obvious. This kind of result, where IRT ends up being somewhat sceptical but not implausibly so, has been a theme of many different cases throughout the book.

9.9 Epistemic Weakness

The cases where cut-elimination fails raise a problem for the way that Fantl and McGrath spell out their version of IRT. Here is a principle they rely on in motivating IRT.

> When you know a proposition p, no weaknesses in your epistemic position with respect to p—no weaknesses, that is, in your standing on any truth-relevant dimension with respect to p—stand in the way of p justifying you in having further beliefs. (Fantl and McGrath, 2009: 64)

And a few pages later they offer the following gloss on this principle.

> We offer no analysis of the intuitive notion of 'standing in the way'. But we do think that, when Y does not obtain, the following counterfactual condition is sufficient for a subject's position on some dimension d to be something that stands in the way of Y obtaining: whether Y obtains can vary with variations in the subject's position on d, holding fixed all other factors relevant to whether Y obtains. (Fantl and McGrath, 2009: 67)

This gloss suggests that the difference between knowledge and evidence is something that stands in the way of an inference. The inquirer who knows that nearby *F*s are *G*s, but does not know that somewhat distant *F*s are *G*s, has many things standing in the way of this knowledge. One of them is, according to this test, that her evidence does not include that all nearby *F*s are *G*s. Yet this is something she knows. So a weakness in her epistemic position with respect to the nature of nearby *F*s, that it is merely evidence and not knowledge, stands in the way of it justifying further beliefs.

The same thing will be true in the monotonic cases of cut-elimination failure. The thinker whose evidence includes $\Gamma \cup \Delta$, and whose inferential knowledge includes A, cannot infer B. But if they had A as evidence, and not merely as knowledge, then they could infer B. So the weakness in their epistemic position, the gap between evidence and knowledge, stands in the way of something.

I didn't endorse the principle of Fantl and McGrath's quoted above, but I did endorse very similar principles, and one might wonder whether they are subject to the same criticism. The main principle I endorsed was that if one knows that p, one is immune from criticism for using p on the grounds that p might be false, or is too risky to use. Equivalently, if the use of p in an inference is defective, but p is known, the explanation of why it is defective cannot be that p is too risky. But now won't the same problem arise? Our inquirer in the monotonic cut-elimination example can't use A in reasoning to B. If A was part of their evidence, then it wouldn't be risky, and they would be able to use it. So the risk is part of what makes the use of it mistaken.

I reject the very last step in that criticism. The fact that something is wrong, and that it wouldn't have been wrong if X, does not mean the non-obtaining of X is part of the ground, or explanation, for why it is wrong. If I break a law, then what I do is illegal. Had the law in question

been struck down by a constitutional court, then my action wouldn't have been illegal. Similarly, if the law had been repealed, my action would not have been illegal. But that doesn't imply that the ground or explanation of the illegality of my action is the court's not striking the law down, or the later legislature not repealing the law. That is to put too much into the notion of ground or explanation. No, what makes the act illegal is that a particular piece of legislation was passed, and this act violates it. This explanation is defeasible—it would be defeated if a court or later legislature had stepped in—but it is nonetheless complete.

The same thing is true in the case of knowledge and evidence. Imagine an inquirer who observes all the *F*s within 3 miles being *G*, and infers both that all the *F*s within 4 miles are *G*, and, therefore, that all the *F*s within 5 miles are *G*. The intermediate step is, in a sense, risky. And the final step is bad. And the final step wouldn't have been bad if the intermediate step hadn't been risky. But it's not the riskiness that makes the second inference bad. No, what makes the second inference bad is that it violates Weisberg's No Feedback principle. That's what the reasoner can be criticised for, not for taking an epistemic risk.

There are two differences then between the core principle I rely on—using reasons that are known provides immunity to criticism for taking epistemic risks—and the principle Fantl and McGrath rely on. I use a concept of epistemic risk where they use a concept of strength of epistemic position. I don't think these are quite the same thing, but they are clearly similar. But the bigger difference is that they endorse a counterfactual gloss of their principle, and I reject any such counterfactual gloss. I don't say that the person who uses known p is immune to all criticisms that would have been vitiated had p been less risky. I just say that the risk can't be the ground of the criticism; something else must be. In some cases, including this one, that 'something else' might be correlated with risk. But it must be the explanation.

Of course, this difference between my version of IRT and Fantl and McGrath's is tiny compared to how much our theories have in common. And indeed, it's tiny compared to how much my theory simply borrows from theirs. But it's helpful I think to highlight the differences to understand the choice points within versions of IRT.