Interests, Evidence and Games

Brian Weatherson

Pragmatic encroachment theories have a problem with evidence. On the one hand, the arguments that knowledge is interest-relative look like they will generalise to show that evidence too is interest-relative. On the other hand, our best story of how interests affect knowledge presupposes an interest-invariant notion of evidence.[1]

The aim of this paper is to sketch a theory of evidence that is interest-relative, but which allows that 'best story' to go through with minimal changes. The core idea is that the evidence someone has is just what evidence a radical interpreter says they have. And a radical interpreter is playing a kind of game with the person they are interpreting. The cases that pose problems for pragmatic encroachment theorists generate fascinating games between the interpreter and the interpretee. They are games with multiple equilibria. To resolve them we need to detour into the theory of equilibrium selection. I'll argue that the theory we need is the theory of **risk-dominant equilibria**. That theory will tell us how the interpreter will play the game, which in turn will tell us what evidence the person has. The evidence will be interest-relative, because what the equilibrium of the game is will be interest-relative. But it will not undermine the story we tell about how interests usually affect knowledge.

1 Encroachment, Reduction and Explanation

I will start with an argument for a familiar disjunctive conclusion: either knowledge is interest-relative, or scepticism is true. The argument will resemble arguments to the same disjunctive conclusion in Hawthorne (2004) and Fantl and McGrath (2009). Indeed, it is inspired by those discussions. But it uses less controversial premises than previous versions.

The argument starts by considering a game, one I'll call the red-blue game. Here are the rules of the game.

1. Two sentences will be written on the board, one in red, one in blue.
2. The player will make two choices.
3. First, they will pick a colour, red or blue.
4. Second, they say whether the sentence in that colour is true or false.

5. If they are right, they win. If not, they lose.
6. If they win, they get $50, and if they lose, they get nothing.

Our player is Parveen. She is an epistemologist who works on pragmatic encroachment, and (as will become important in a minute), she has frequently cited both *Knowledge and Lotteries* (Hawthorne, 2004), and *Knowledge and Practical Interests* (Stanley, 2005). She knows the rules of the game, and no other relevant facts about the game. When the game starts, the following two sentences are written on the board, the first in red, the second in blue.

1. Two plus two equals four.
2. *Knowledge and Lotteries* was published before *Knowledge and Practical Interests*.

Intuitively, there is a unique rational play in this game: Red-True. That is, Parveen announces that she will evaluate the truth value of the red sentence, and then announce that it's true. That's a sure $50.

On the other hand, in normal circumstances, we would say that Parveen does know that *Knowledge and Lotteries* was published before *Knowledge and Practical Interests*. After all, she has looked up their publication dates many times in checking over her papers.

There is a puzzle in reconciling these intuitions. The pragmatic encroachment theorist has a solution to these puzzles. In normal circumstances, Parveen does know that *Knowledge and Lotteries* was published before *Knowledge and Practical Interests*. But these are not normal circumstances. Right now, it matters whether her reason to believe that *Knowledge and Lotteries* was published before *Knowledge and Practical Interests* is as strong as her reason to believe that two plus two equals four. And (unless something very weird is happening), that isn't true for Parveen. So she knows that red-true will win, she doesn't know any other play will win, so she should play Red-True.

If we reject pragmatic encroachment, and we are not sceptics, we should say that Parveen does know that *Knowledge and Lotteries* was published before *Knowledge and Practical Interests*. And then it is a mystery why playing Red-True is more rational than playing Blue-True. After all, Parveen knows the rules of the game, and she knows (by hypothesis) the blue sentence is true, so if she can do even basic logical reasoning in a knowledge preserving way, she knows she will get as good a result as possible by playing Blue-True. So it is a bit of a mystery why it would be anything other than maximally rational to play Blue-True.

One way we might try to resolve this mystery is by saying that although Parveen knows that Blue-True will win $50, she super-knows that Red-True will win $50. What do we mean here by *super knowledge*? Think of this as a placeholder for certainty, or knowledge that one knows, or anything other epistemic state that you think might be relevant to her practical decision making. Perhaps the fact that she super-knows what two plus two is, but doesn't super-know when the epistemology books were published, could be the explanation for why Red-True is the unique rational play.[2]

But no such explanation can work, because Parveen doesn't super-know that playing Red-True will win $50. She super-knows that two plus two is four. But we have not assumed that she super-knows the rules of the game. So she doesn't super-know that Red-True will win, she just knows it. And she also, by hypothesis, knows that Blue-True will win. So looking at any kind of super-knowledge can't break the intuitive asymmetry between Red-True and Blue-True.

Put another way, if Parveen knows that *Knowledge and Lotteries* was published before *Knowledge and Practical Interests*, then she knows that she is playing the following game.

1. Two sentences will be written on the board, one in red, one in blue.
2. The player chooses to play either Blue-True, Blue-False, Red-True, or Red-False.
3. If they play Blue-True, they win $50.
4. If they play Blue-False, they win nothing.
5. If they play Red-True, they win $50 if the red sentence is true, and nothing otherwise.
6. If they play Red-False, they win $50 if the red sentence is false, and nothing otherwise.

And is is rational to play Blue-True in that game. (It might also be rational to pay Red-True depending on what the red sentence is, but it is always rational to play Blue-True.) Yet it is not rational to play Blue-True in the original game. So Parveen does not know, when she plays the original game, that *Knowledge and Lotteries* was published before *Knowledge and Practical Interests*.

So to avoid pragmatic encroachment here we must deny that Parveen ever knew that *Knowledge and Lotteries* was published before *Knowledge and*

---

[2]That we need some kind of super-knowledge for action, and not mere knowledge, is a popular, and natural, explanation of the case. For versions of this explanation, obviously with more details than I've given here, see for example Jessica Brown (2008) and Jennifer Lackey (2010).

Practical Interests. On its own, that's not a sceptical conclusion: lots of people don't know that. But once we go down that path, it looks like not much knowledge will be left. After all, we can repeat the game with any number of different things in the place of the blue sentence. If we adopt the constraint that Parveen only knows *p*, right now, if it is rationally permissible for her to play Blue-True when *p* is the blue sentence, no matter what the red sentence is, then either we have to say very unintuitive things about rational plays of the game, or we have to say she knows very little.

So we've got the conclusion that either pragmatic encroachment is true, or scepticism is true. Since I'm not a sceptic, I'm happy to conclude that pragmatic encroachment is true. But note that we've done this without any reference to high stakes situations. The stakes in Parveen's game are just $50. That's not nothing, but it's not 'high stakes' in the way that phrase is normally used.

The version of pragmatic encroachment we get is that what matters for knowledge are not the stakes involved in any bet on *p*, but the odds.[3] Parveen loses knowledge because she is being asked, in effect, to make a super long odds bet on a fact about publication schedules. She is in no position to rationally make a bet at those odds. So she doesn't know the fact about publication schedules.

And that's the general principle: agents only know a proposition if they are in a position to rationally bet on that proposition at the odds currently being offered to them. In practice, high stakes situations tend to feature bets at long odds, so in practice much knowledge dissipates in high stakes cases. But the explanation of the dissipation is the odds the agent faces, not the stakes.

More precisely, I endorse these principles as constraints on knowledge:

- If the agent knows that *p*, then for any question they have an interest in, the answer to that question is identical to the answer to that question conditional on *p*.

---

[3] Jessica Brown (2008, 176) shows that pragmatic encroachment theories that rely just on the stakes involved are subject to serious counterexample. Katherine Rubin (2015) argues that if we have a 'global' version of pragmatic encroachment, where all our epistemic notions are interest-relative, then it is implausible that it is the stakes the subject faces that matter for knowledge. Since I'm defending such a global version of pragmatic encroachment, Rubin's arguments show that it is important that I'm relying on odds, not stakes. Baron Reed (2014) argues that if it is stakes alone that matter to pragmatic encroachment, then agents who the pragmatic encroachment theorist takes to be perfectly rational would be subject to a Dutch Book.

- When an agent is considering the choice between two options, the question of which option has a higher expected utility given their evidence is a question they have an interest in.

Those principles are meant to not merely be extensionally adequate. They are meant to explain why agents lose knowledge when considering some sets of options, like in the Red-Blue game. In some sense, they are meant to be part of reductive explanations. These reductive explanations take as primitive inputs facts about the agent's evidence, and facts about evidential probability. I'm going to set aside worries about the metaphysics of evidential probability, and just focus on evidence. Because it turns out that there is a real problem in getting a plausible theory of evidence that can function as an input to that reductive explanation.

2   The Problems with Evidence

Go back to the red-blue game. Consider a version of the game where:

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

I'm going to argue that there are cases where the only rational play is Red-True, but the blue sentence is something we want to say that, ordinarily, the subject knows. And I'll argue that this is a problem for the kind of reductive explanation I just sketched. If pragmatic effects matter to what the evidence is, we can't take the evidence as a fixed input into an explanation of how and when pragmatic effects matter.

Let's have Parveen play the game again. She's going to be playing the game in a restaurant, one in Ann Arbor where she lives. Just before the game starts, she notices an old friend, Rahul, across the room. Rahul is someone she knows well, and can ordinarily recognise, but she had no idea he was in town. She thought Rahul was living in Italy. Still, we would ordinarily say that she now knows Rahul is in the restaurant; indeed that he is in the restaurant. It would be perfectly acceptable for her to say to someone else, "I saw Rahul here", for example. Now the game starts.

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

Now we have a problem. On the one hand, there is only one rational play here: Red-True. If you haven't seen someone for a long time, then you can't be completely certain it's them when you spot them across a restaurant. It would be foolish to be as confident that it's Rahul as that two and two make four. It looks like this is a case where pragmatic effects defeat knowledge.

On the other hand, our story for why Parveen loses knowledge here has run into problems. I wanted to tell a story roughly like the following. She can't play Blue-True when the probability of the blue sentence, given her evidence, is less than the probability of the red sentence, given her evidence. That explanation can only go through if the blue sentence is itself not part of her evidence, since the probability of anything given itself is one. So we need a story about how it is that it is not part of Parveen's evidence that Rahul is not in the restaurant.

That story can't be the one that presupposes facts about what is in Parveen's evidence. So it can't use facts about the probability of some proposition given her evidence; at least not in any simple way. If we can independently identify Parveen's evidence, then we can go back to using evidential probability. But until we've done that, we're stuck.

There are two options here that seem possible for the pragmatic encroachment theorist, but not particularly attractive.

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

Another option is to deny that a reductive explanation is needed here. Perhaps pragmatic effects, like the particular sentences that are chosen for this instance of the Red-Blue game, mean that Parveen's evidence no longer includes facts about Rahul, but this isn't something we can give a reductive account of. We shouldn't assume that everything will have a simple reductive explanation, so this isn't so bad in theory. The problem in practice is that without a reductive explanation, we don't have a predictive theory of when pragmatic effects matter. And that seems to be a bad thing. For instance, the following theory is completely consistent with Parveen's case as described.

1. E=K; i.e., one's evidence is all and only what one knows.
2. Someone does not know p if the evidential probability of p is not close enough to one for current purposes.
3. Since it is part of Parveen's evidence that Rahul is in the restaurant, the probability that he is there is one, so it is close enough to one for current purposes.
4. So this is not a case where pragmatic effects change what she knows.

That theory seems to me to be badly mistaken, since it goes on to predict that it is rationally permissible to play Blue-True. But we need a pragmatic account that says that it is mistaken, and says something about which alternative situations would not threaten Parveen's knowledge. We don't yet, as far as I can see, have such an account. The aim of the rest of this paper is to provide one.[4]

3 A Simple, but Unsatisfying, Solution

Let's take a step back and look at the puzzle more abstractly. We have an agent S, who has some option O, and it really matters whether or not the value of O, i.e., $V(O)$ is at least x. It is uncontroversial that the agent's evidence includes some background K, and controversial whether it includes some contested proposition p. It is also uncontroversial that $V(O|p) \geq x$, and we're assuming that for any proposition q that is in the agent's evidence, $V(O|q) = V(O)$. That is, we're assuming the relevant values are conditional on evidence. We can capture that last assumption with one big assumption that probably isn't true, but is a harmless idealisation for these purposes. Say there is a prior value function V^-, with a similar metaphysical status to the mythical, mystical prior probability function. Then for any choice C, $V(C) = V^-(C|E)$, where E is the evidence the agent has.

Now we're in a position to state a simple, but unsatisfying, solution. Let p be the proposition that the agent might or might not know, and the question of whether $V(O) \geq x$ be the only salient one that p is relevant to. Then the agent knows p only if the following is true:

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

---

[4]You can read this paper as a reply to the challenge posed by Ichikawa et al. (2012). They note that there are challenges facing the pragmatic encroachment theorist whether they make evidence interest-relative, or interest-invariant. I'm going to show how to have an interest-relative theory of evidence, and keep what was desirable about pragmatic encroachment theories.

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

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

Great! Well, great except for two monumental problems. The first problem is that what we've said here really only helps with very simple cases, where there is a single decision problem that a single contested proposition is relevant to. We need some way to generalise the case to less constrained situations. The second (and bigger) problem is that the solution is completely ad hoc. Why should we use the arithmetic mean of these two things rather than any other formula that would have implied the intuitively correct result in the Parveen-Rahul case? Pragmatic encroachment starts with a very elegant, very intuitive, principle: you only know the things you can reasonable take to be settled for the purposes of current deliberation. And that deliberation should be driven by the aim of maximising expected utility. It does not look like any such elegant, intuitive, principles will lead to some theorem about averaging out the value of an option with and without new evidence.

Happily, the two problems have a common solution. But the solution requires a detour into some technical work. It's time for some game theory.

4 Gamifying the Problem

We can usefully think of some philosophical problems as games, and hence subjects for study using game theoretic techniques. This is especially when the problems involve interactions of rational agents. Here, for example, is the game table for Newcomb's problem, with the human who is usually the focus of the problem as Row, and the demon as Column.[5]

---

[5]In these games, Row chooses a row, and Column chooses a column, and that determines the cell that is the outcome of the game. The cells include two numbers. The first is Row's payout, and the second is Column's. The games are non-competitive; the players are simply trying to maximise their own returns, not maximise the difference between their return and the other player's return.

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

This game has a unique Nash equilbrium; the bottom right corner.[6] And that's one way of motivating the view that (a) the game is possible, and (b) the rational move for the human is to choose two boxes.

Let's look at a more complicated game. I'll call it The Interpretation Game. The game has two players. Just like in Newcomb's problem, one of them is a human, the other is a philosophical invention. But in this case the invention is not a demon, but The Radical Interpreter.[7] To know the payouts for the players, we need to know their value function. More colloquially, we need to know their goals.

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

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

Let's make all that (marginally) more concrete. Human is offered a bet on p. If the bet wins, it wins 1 util; if the bet loses, it loses 100 utils. Human's only choice is to Take or Decline the bet. The proposition p, the subject of the bet, is like the claim that Rahul is in the restaurant. It is something that is arguably part of Human's evidence. Unfortunately, it is also arguable that it is not part of Human's evidence. We will let K be the rest of Human's evidence

---

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

[7]The Radical Interpreter feels like they should be a humanesque character in *Alice in Wonderland* or *The Phantom Tollbooth*, but for now they are resolutely abstract.

(apart from p, and things entailed by $K \cup \{p\}$), and stipulate that $\Pr(p|K) = 0.9$. Each party now faces a choice.

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

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

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

We have, in effect, already covered The Radical Interpreter's payouts. They win in the top-left and lower-right quadrants, and lose otherwise. Human's payouts are only a little trickier. In the bottom row, they are guaranteed 0, since the bet is declined. In the top-left, the bet is a sure winner; their evidence entails it wins. So they get a payout of 1. In the top-right, the bet wins with probability 0.9, so the expected return[8] of taking it is $1 \times 0.9 - 100 \times 0.1 = -9.1$.

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

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

---

[8]I am making a large, if orthodox, assumption here: that the payouts that we use for equilibrium analysis should be expected returns, not actual returns. I think that's the right thing to do, since it is usually impossible to say what the actual return of a game is. Even when we say that the payout is a certain number of dollars, we are really saying that the return is a certain kind of gamble. Maybe the value of the currency will depreciate quickly, and the dollars are not that valuable. Maybe the revolution will come and wealth will be a liability. Almost all games have probabilistic payouts, and this game is no different.

The mathematical result that there are two equilibria to this game should not come as a surprise. In discussing games like this earlier, we said that general principles connecting evidence, knowledge and action are not predictive; they are consistent both with p being part of the evidence, and with it not being part of the evidence. The general principles we had stated rule out, in effect, non-equilibrium solutions to games like this one. But they are not predictive in cases where there are multiple equilibria.

To make more progress, we need to turn to more contested areas of game theory. In particular, we need to look at some work on equilibrium choice. We'll introduce this material via a game that is inspired by an example of Rousseau's.

5 Equilibrium Selection Principles

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

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

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

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

The first four clauses say that the game has two (strict) Nash equilibria: *Aa* and *Bb*. The fifth and sixth clauses say that the *Aa* equilibria is **Pareto-optimal**: no one prefers the other equilibria to it. In fact it says something a bit stronger: one of the players strictly prefers the *Aa* equilibria, and the other player does not prefer *Bb*. The seventh and eighth clauses say that the *Bb* equilibria is **risk-optimal**. Risk-optimality is a somewhat complicated notion in general; see Harsanyi and Selten (1988) for more details. But for our purposes, we can focus on a simple characterisation of it. Neither player

would prefer playing *A/a* to playing *B/b* if they thought it was a coin flip which equilibrium the other player was aiming for.

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

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

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

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

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

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

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

Embed this game in what they call a *global game*. We'll start the game with each player knowing just that they will play a game with the following payout table, with x to be selected at random from a flat distribution over $[-1, 5]$.

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

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

In particular, they prove that iterated deletion of strictly dominated strategies (almost) removes all but one strategy pair.[9] Each player will play *A/a* if the signal is greater than 2, and *B/b* otherwise.[10] Surprisingly, this shows that players should play the risk-optimal strategy even when they know the other strategy is Pareto-optimal. When a player gets a signal in $(2, 3.75)$, then they know that $x < 4$, so *Bb* is the Pareto-optimal equilibrium. But the logic of the global game suggests the risk-dominant equilibrium is what to play.

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

I conclude from that that risk-dominant choices are rational in these games. There is a limit assumption involved here; what's true for games with arbitrarily small margins of error is true for games with no margin of error. (We'll come back to that assumption below.) And since The Radical Interpreter is

---

[9]A sketch of the proof is in Appendix One.

[10]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.

rational, they will play the strategy that is not eliminated by deleting dominant strategies. That is, they will play the risk-dominant strategy.

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

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

In short, this model gives us a way to keep what was good about the pragmatic encroachment theory, while also allowing that evidence can be interest-relative. It does require a slightly more complex theory of rationality than we had previously used. Rather than just say that agents maximise evidential expected utility, we have to say that they play risk-dominant strategies in coordination games. But it turns out that this is little more than saying 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.

6 Objections and Replies

We'll end the body of the paper with some objections that might be raised to this model. And then the appendix will contain proofs of a couple of the formal claims.

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

Reply: If pushed, I would defend the use limit assumptions like this to resolve hard cases like Stag Hunt. But I don't need that assumption here, What we really need is that Parveen doesn't know precisely the probability of Rahul being in the restaurant given the rest of her evidence. Given that evidence is not luminous, as Williamson (2000) shows, this is a reasonable assumption.

So the margin of error assumption that Carlsson and van Damme make is not, in our case, an assumption that merely makes the math easier; it is built into the case.

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

Reply: It turns out that Carlsson and van Damme's result doesn't require that both parties are ignorant of the precise values of the payoffs. As long as one party doesn't know the exact value of the payoff, the argument goes through. I prove this in Appendix Two.

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

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

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

Reply: Sadly, this is true. I would like to have a simpler model, but I don't know how to create one. The argument I gave earlier that our simple principles underdetermine what to say in cases like Parveen and Rahul's seems fairly compelling. So more complexity will be needed, one way or another. I think paying this price in complexity is worth it overall, but I can see how some people might think otherwise.

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

Reply: This is all true, and it shows that the view I'm putting forward is incompatible with some programs in epistemology. In particular, it is incompatible with E=K, since the what it takes to be evidence on this story is

slightly different from what it takes to be knowledge. I don't think E=K is so intuitively obvious that this refutes the theory, but it is potentially a cost that I have to give it up.

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

Reply: This is somewhat true. There are other ways of embedding the game involving Human and The Radical Interpreter in global games that lead to different outcomes. They are usually somewhat artificial; e.g., by having the signal be systematically biased in one way. But what really matters is the game where the error in Human's knowledge of the payoffs is determined by their actual epistemic limitations. I think that will lead to something like the model we have here. But it is possible that the final result will differ a bit from what I have here, or (more likely) have some indeterminacy about just how interests interact with evidence and knowledge. The precise details are ultimately less important to me than whether we can provide a motivated story of how interests affect knowledge and evidence that does not presuppose we know what the agent's evidence is. And the method I've outlined here shows that we can do that, even if we end up tinkering a bit with the details.

References

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

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

Fantl, Jeremy and McGrath, Matthew. 2009. *Knowledge in an Uncertain World*. Oxford: Oxford University Press.

Harsanyi, John C. and Selten, Reinhard. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

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

Ichikawa, Jonathan Jenkins, Jarvis, Benjamin, and Rubin, Katherine. 2012. "Pragmatic Encroachment and Belief-Desire Psychology." *Analytic Philosophy* 53:327–343, doi:10.1111/j.2153-960X.2012.00564.x.

Lackey, Jennifer. 2010. "Acting on Knowledge." *Philosophical Perspectives* 24:361–382.

Reed, Baron. 2014. "Practical Matters Do Not Affect Whether You Know." In Matthias Steup, John Turri, and Ernest Sosa (eds.), *Contemporary Debates in Epistemology*, 95–106. Chicester: Wiley-Blackwell, 2nd edition.

Rousseau, Jean-Jacques. 1913. *Social Contract & Discourses*. New York: J. M. Dent & Sons. Translated by G. D. H. Cole.

Rubin, Katherine. 2015. "Total Pragmatic Encroachment and Epistemic Permissiveness." *Pacific Philosophical Quarterly* 96:12–38, doi:10.1111/papq.12060.

Schwitzgebel, Eric. 2008. "The Unreliability of Naive Introspection." *Philosophical Teview* 117:245–273.

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

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.

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

Appendix One: Carlsson and van Damme's Game

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

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

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

Before getting the signal, the players must choose a strategy. A strategy is a function from signals to choices. Since the higher the signal is, the better it is to play B, we can equate strategies with 'tipping points', where the player plays B if the signal is above the tipping point, and A below the tipping point. Strictly speaking, a tipping point will pick out not a strategy but an

equivalence class of strategies, which differ in how they act if the signal is the tipping point. But since that happens with probability 0, the strategies in the equivalence class have the same expected return, and so we won't aim to distinguish them.

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

Call the tipping points for Row and Column respectively T_R and T_C. Since the game is symmetric, we'll just have to show that in conditions of common knowledge of rationality, $T_R = 2$. It follows by symmetry that $T_C = 2$ as well. And the only rule we'll use is iterated deletion of strictly dominated strategies. That is, we'll assume players won't play strategies where another strategy does better no matter what the opponent chooses, and they won't play strategies where another strategy does better provided the other player does not play a dominated strategy, and they won't play strategies where another strategy does better provided the other player does not play a strategy ruled out by these first two conditions, and so on.

The return to a strategy is uncertain, even given the other player's strategy. But given the strategies of each player, we can work out an expected return for each player. And that's what we'll assume is the return to a strategy pair.

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

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

Now comes the important step. Assume that we know that $T_C \leq y > 2$. Now consider the expected return of playing A given various values for $s_R > 2$. Given that the lower T_C is, the higher the expected return is of playing A, we'll just work on the simple case where $T_C = y$, realizing that this is an upper

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

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

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

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

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

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

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

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

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

Appendix Two: The Modified Game

The aim of this section is to prove something that Carllson and van Damme did not prove, namely that the analysis of the previous appendix goes through with very little change if one party gets a perfect signal, while the other gets a noisy signal. That is, we're going to consider the game that is just like the game of appendix one, but it is common knowledge that the signal Column gets, s_C, equals x.

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

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

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

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

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

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

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

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