Interests, Evidence and Games

Brian Weatherson, July, 2017

Summary

- The best model of pragmatic encroachment assumes that we are given a notion of evidence.
- But the best arguments for pragmatic encroachment generalise to show that evidence is interest-relative.
- So we need a model that solves simultaneously for the variables *What is the evidence?* and *What is known?*, given the physical situation and the interests.
- The motivating idea here comes from radical interpretation: the evidence is what the radical interpreter would interpret the evidence to be.
- I then use some tools from game theory to fill out how this interpretation might get done, and show how to allow the 'best model' to work while evidence itself is interest-relative.

The Red-Blue Game

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

Parveen knows the rules and nothing else, and plays the game with these two sentences.

Red Two Plus Two Equals Four.
Blue *Knowledge and Lotteries* was published before *Knowledge and Practical Interests*.

1. The only rational play is Red-True.
2. If Parveen knew **Blue**, then Blue-True would be rational.
3. So Parveen (while playing the game) doesn't know **Blue**.
4. This generalises to scepticism unless knowledge is interest-relative.
5. So scepticism is true, or knowledge is interest-relative

If super-knowledge is required for action, then we can't explain why Red-True is rationally required. If Parveen knows **Blue**, we can't distinguish this game from the following one where Red-True is not rationally required.

- A single sentence, in this case *Two Plus Two Equals Four* is written in red.
- There are four choices: Red-True, Red-False, Blue or Green.
- If Red-True, get $50 if the sentence is true, $0 otherwise. Red-False is other way around.
- If Blue, get $50.
- If Green, get nothing.
- Player knows these rules, the identity of the sentence, and nothing else.

We get interest-relativity (or pragmatic encroachment if you prefer), but it's odds that matter, not stakes. After all, $50 is low stakes.

I endorse these principles, and take them to be both extensionally adequate and explanatory.

- 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.
- 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.

The Problem 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.

Here's the problem. Red-True is still the only rational play. But we can't explain that in terms of evidential probabilities, because what's at issue is whether **Blue** is part of the evdience. Here are some ways out.

1. Make evidence psychological, and infallible, so that Red-True is not the unique rational choice.
2. Deny that the story of the last section is explanatory.

These seem bad, so I'm going to look for a different approach.

A Simple, but Unsatisfying, Solution

Start with this abstract version of the case.

- We have agent S with option O, and value function V.
- It really matters whether $V(O) \geq x$, and nothing else about O matters.
- It is uncontroversial that the evidence includes K.
- It is controversial whether the evidence includes p, and p is relevant to O and nothing else.
- And assume (controversially) that there is a 'prior' value function $V^{\vdash}$, so for any choice C, $V(C) = V^{\vdash}(C|E)$, where E is the evidence the agent has.

Hypothesis: The agent knows p only if this obtains:

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

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

- Good news: This gets the above cases right.
- Bad news: This is absurdly ad hoc.
- Worse news: We don't even have a way to generalise it to cases with multiple options, multiple things p is relevant to, etc.

It's time for some game theory.

Gamifying Problems

We can usefully think of several philosophical problems as games. Here, for example, is the game table for Newcomb's problem, with Row as Human, and Column as Demon. Note that in all 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.

	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. 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.

The Interpretation Game

The game has two players: Human and The Radical Interpreter. Our first job is to set their value function, via thinking about 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 The Radical Interpreter.

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.

- Human is offered a choice to Take or Decline a bet on p.
- If the bet wins, it wins 1 util; if the bet loses, it loses 100 utils.
- If p is known, it is part of Human's evidence, but it is not obviously known.
- Let K be the rest of Human's evidence (apart from p, and things entailed by $K \cup \{p\}$), and stipulate that $\Pr(p|K) = 0.9$.
- 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. 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

A quick explanation of Human's payouts. In the bottom row, they are guaranteed 0, since the bet is declined. In the top-left, the bet is a sure winner; their evidence entails it wins. So they get a payout of 1. In the top-right, the bet wins with probability 0.9, so the expected return of taking it is $1 \times 0.9 - 100 \times 0.1 = -9.1$.

There are two Nash equilibria for the game, top-left and bottom-right. To make more progress, we need to go beyond equilibrium analysis, to principles that select among equilibria. And we'll do it via an example of Rousseau's.

Equilibrium Selection Principles

Here's a maximally abstract version of a 2x2 game.

	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 eight properties listed to the right:

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

- Clauses 1-4 imply *Aa* and *Bb* are the only pure strict Nash equilibria.
- Clauses 5 and 6 say that the *Aa* equilibria is **Pareto-optimal**: no one prefers the other equilibria to it.
- Clauses 7 and 8 say that the *Bb* equilibria is **risk-optimal**. Roughly, each would prefer it if they thought the other would flip a coin between the equilibrium choices.

I claim in these cases, the risk-optimal equilibrium is the rational one. Games satisfying these eight inequalities are sometimes called *Stag Hunt* games. 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 assumes here, rightly I think, that the (short-term) rational play is the uncooperative one.

There is a recent argument to the same conclusion by Hans Carlsson and Eric van Damme. Assume that the payoffs are not perfectly known - some of them include small (and symmetric) error bars. Then iterated strict dominance reasoning implies that the risk-optimal strategy is uniquely rational. (This is not at all obvious, but it would be a big detour to prove it here. I'll say more in Q & A if people are interested.)

Back to Interpretation

The risk-dominant equilibria in the Interpretation Game we had is that $p \notin E$, and Human declines the bet. I think this is rational for both players, and since The Radical Interpreter does what is rational, it is true that $p \notin E$. And that's the general case; the evidence is what The Radical Interpreter says the evidence is, assuming that they choose risk-dominant equilibria. This view has a number of striking features:

- It is interest-relative.
- It is systematic; we can apply it to more than simple cases, though computing risk-dominant equilibria gets non-trivial after a while.
- It is not (completely) ad hoc; it follows from independent principles about strategy choice.
- It is reductive, or at least as reductive as the original interest-relative story.

So the interest-relative theorist can keep their general story of how interests matter, and allow evidence to be interest-relative too.