The Partition Assumption

Abram Demski

some rough motivation

- Questioning i/o assumptions ("agent boundaries")
- Discussions about representation theorems, EG Savage vs. J-B, tend to foreground actions (output).
- I want to foreground observations (input). In order to do this, we have to look at how people apply these theories.

Again & again, we see Partitional Evidence.

- Aumann's Common knowledge
- Aumann's Agreement Theorem

} some work relaxing PE here

- Solomonoff prior & AIXI
- Information Theory

} some work relaxing PE here

- Cartesian Frames

↳ PE is part of the def. of "observation" here, but easy to explore variations

Partitional Evidence Assumption:

Let $E(w)$ be the set of worlds encoding the "evidence" at world w;
IE the worlds the agent thinks could be.

This is assumed to be a partition function:

1. Reflexivity*: $w \in E(w)$
2. Transitivity*: $x \in E(y) \;\&\; y \in E(z) \rightarrow x \in E(z)$
3. Symmetry*: $x \in E(y) \leftrightarrow y \in E(x)$

*: for the relation $a \in E(b)$

All three axioms are questionable.

Reflexivity:

- From outside the agent, we see that $w \notin E(w)$ in corrupt states.
- From inside, this axiom is like $\Box A \to A$, which is problematic via Löb, given Trans.

accessibility relation in Kripke frame	modal logic
Reflexivity	$\Box A \to A$
Transitivity	$\Box A \to \Box\Box A$
Symmetry	$A \to \Box\Diamond A$

Löb's Theorem:

Diagonalization

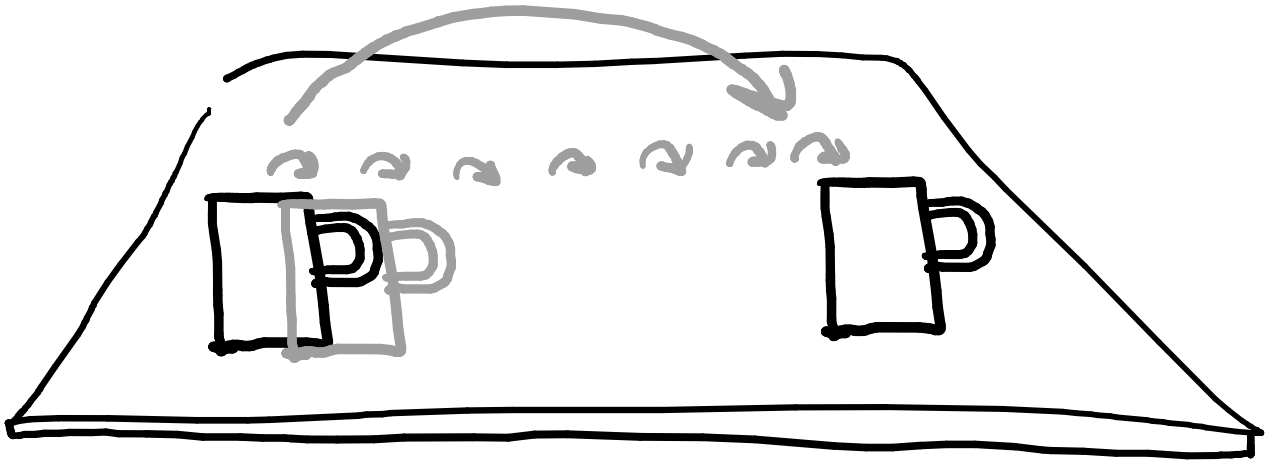
+ $\vdash X \Rightarrow \vdash \Box X$
+ $\vdash \Box X \Rightarrow \vdash \Box\Box X$
+ $\Box(X \to Y) \to (\Box X \to \Box Y)$

$\Box(\Box X \to X) \to \Box X$

(Reflexivity continued)

- We could simply minimize fallibility & analyze things under an assumption that feedback is not corrupt.
- We can also retreat to "phenomenological" evidence: observations whose only referential content is "I have been observed" But, Löb!
- I don't have strong arguments against, but I want to explore a different way.

Transitivity:



- Sufficiently small differences in the location of a coffee cup are imperceptible, but we can chain these together to get large differences.
- This may sound irrelevant to digital machines, but it does become critical in Garrabrant Induction and in Paul's self-knowing probabilities.

Symmetry:

Modesty Argument: (pro-symmetry)

> Many people, while dreaming, are not aware that they are dreaming. Many people, while dreaming, may believe at some point they have woken up, while still being asleep. Clearly there can be no license from "I think I'm awake" to the conclusion that you actually are awake, since a dreaming person could just dream the same thing.

Eliezer: (anti-symmetry)

> Those who dream do not know they dream, but when you are awake, you know you are awake.

(symmetry cont.)

We can also reformulate symmetry:

> **Negative Introspection**: For any world w, if the evidence at w does not entail some proposition P, then the evidence entails that it does not entail P. $\neg\Box P \rightarrow \Box\neg\Box P$

This is equiv. to Transitivity given the other axioms, but not in their absense.

We can think of transitivity as Positive Introspection. $\Box P \rightarrow \Box\Box P$

Why does it matter? (more motivation)

- RL has the wireheading problem.
  - corrupt feedback
  - can't rule out "wirehead hypothesis"
- If we can construct an observation-utility fn, we have the human manip. problem.
- Any restricted notion of feedback has something similar.

→ Stable Pointers to Value sequence (me)
→ Pointers Problem (John)
→ Observation-Utility (D.D.)
→ Corrupt Reward Channel (Tom E.)

Learning Normativity Agenda

→ Feedback on the whole event space
(contradicts Partition Assumption)
↳ feedback on the structure of good hypotheses can address inner opt. in principle

→ Feedback "always uncertain"
(can always be corrupt)
↳ plausibly, denies Reflexivity

Why do we think we need Partitional Evidence?

The Problem of Filtered Evidence.

Meta-Update Formalization:

Updating on $E(\omega)$ should be equivalent to the "meta-update" on the fact that our evidence is $E(\omega)$.

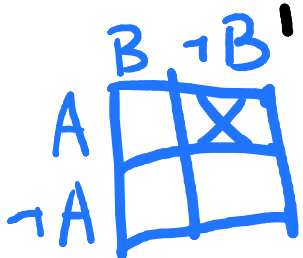
I strongly recommend the discussion from Pearl.
(Probabilistic Reasoning in Intelligent Systems, Chapter 2.)

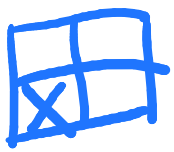
But he advocates meta-update-equivalence.
I will argue that this is computationally infeasible.

Example: Trolling Mathematicians
(An Untrollable Mathematician)

(Scott, 2014)

> Suppose you're interested in A. A troll can drive your credence in A down by finding some true $A \to B$ and proving it to you. Or drive it up with some true $B \to A$.





By iterating this process, the troll can drive belief in A arbitrarily high, then low, then high, back & forth arbitrarily many times.

But: it seems like you should be able to escape this trap by understanding how the troll is giving you filtered evidence.

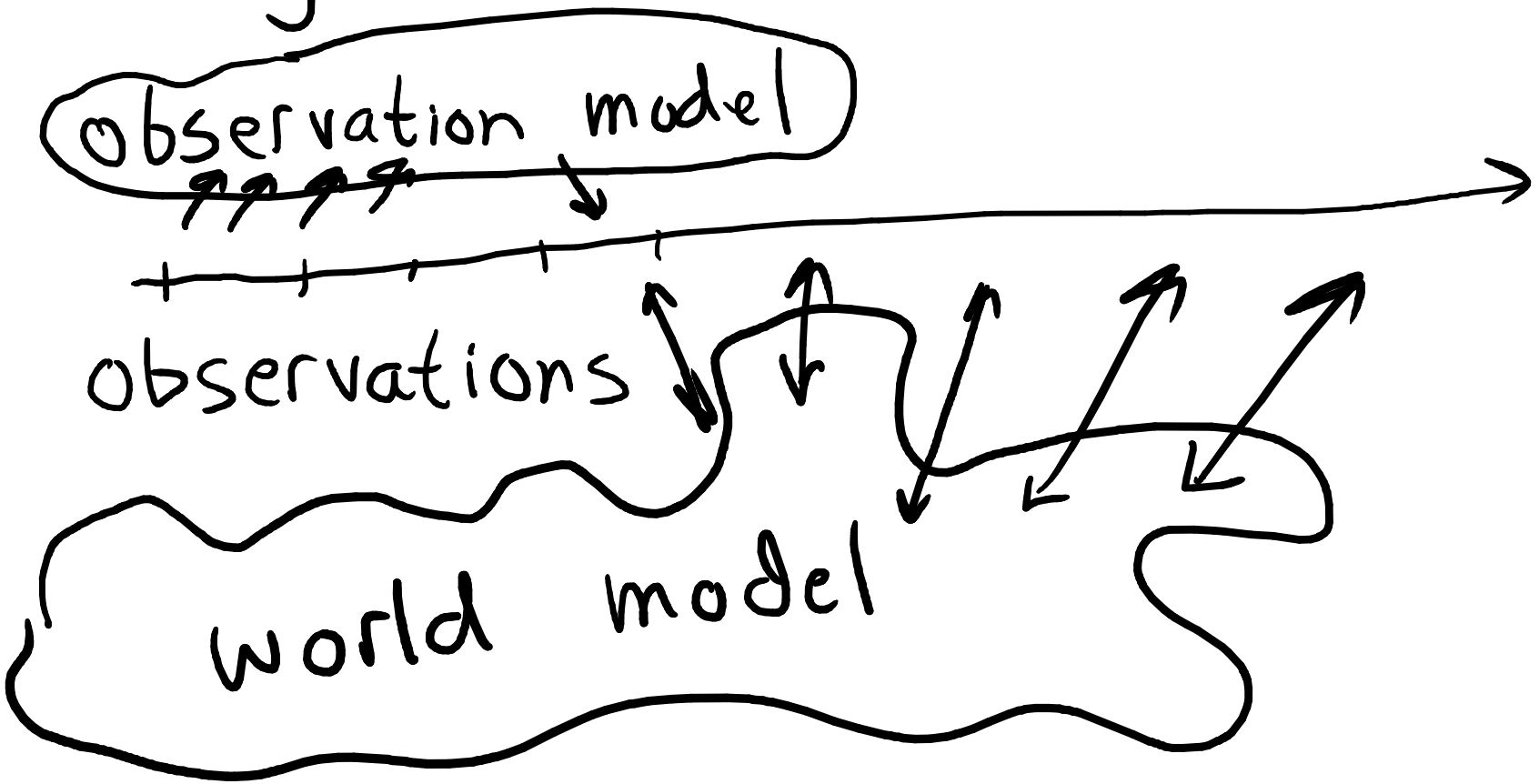
In 2018, Sam came up with such a method.

But Sam's method has a striking flaw: the prior which escapes trolling is really dumb.

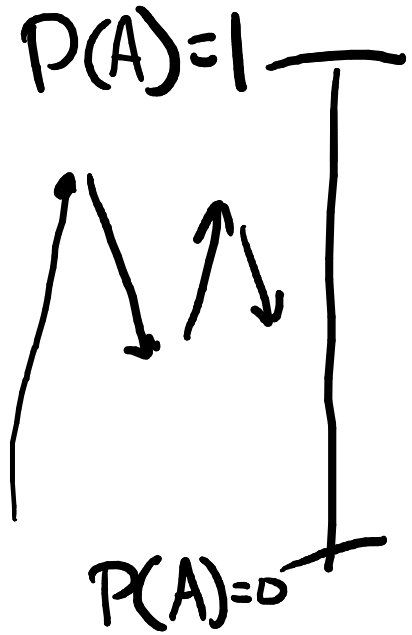
It turns out this is an essential feature of all such solutions.

(Thinking About Filtered Evidence is (very!) Hard, 2020)

The basic problem is: it is relatively computationally easy to predict observations one at a time; but the world model can't keep up with the arbitrarily-far-out predictions even relatively simple observation models can make.



Radical Probabilism (in the guise of Garrabrant Induction) solves the problem in a more computationally feasible way.



If the agent is trolled, ie some belief goes back & forth, traders can profit by buying low & selling high, which dampens the oscillations.

A Way Out of Partitional Evidence?

Radical Probabilism reconceives evidence as anything which shifts probabilities, abandoning even the assumption that there is necessarily an evidence proposition $E(\omega)$.

→ No fixed utility fn; it can be revised like everything else.

→ A model of revising beliefs through philosophy, which seems potentially critical.

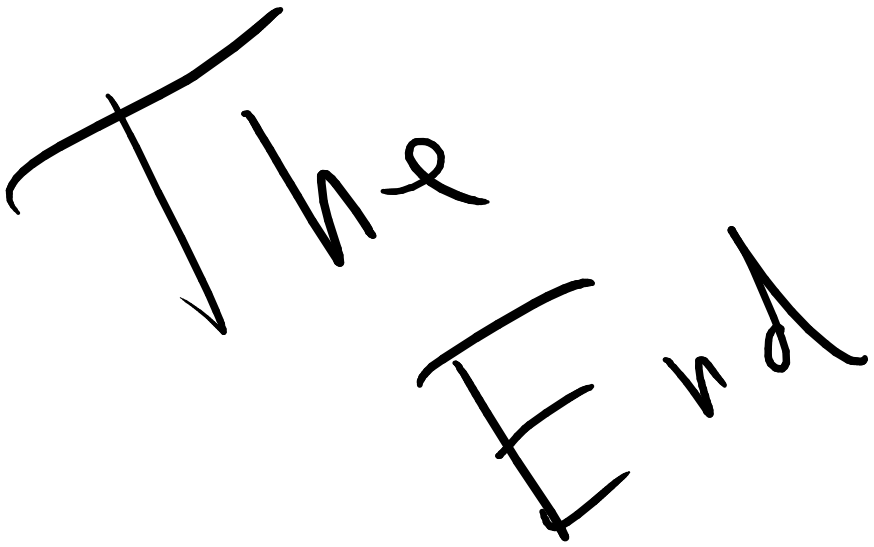
Summary

→ Partitional Evidence underlies many models, even though you don't see it listed in probability/rationality axioms.

→ All three axioms seem questionable.

→ While it technically helps us mitigate filtered evidence, it does so in a way which is inconsistent w/ general learning.

→ Radical Probabilism rejects it & gives us tools which may aid AI Alignment.


The End