Higher-Order Evidence, Accuracy, and Information Loss

Ben Levinstein

Rutgers University

04 December 2016

Accuracy First Epistemology

Laudable properties of credences:

- Informativeness
- Simplicity
- Unification
- Justification
- **Accuracy**

Accuracy is *the* fundamental epistemic good.

- The higher your credence in truths and the lower your credence in falsehoods, the better off you are all epistemic things considered.

Consequentialist: facts about the epistemic *good* (accuracy) explain what's epistemically *right*.

- Epistemic norms have binding force in virtue of helping in the pursuit of *accurate credences*

Accuracy First Epistemology

Laudable properties of credences:

- Informativeness
- Simplicity
- Unification
- Justification
- **Accuracy**

Accuracy is *the* fundamental epistemic good.

- The higher your credence in truths and the lower your credence in falsehoods, the better off you are all epistemic things considered.

Consequentialist: facts about the epistemic *good* (accuracy) explain what's epistemically *right*.

- Epistemic norms have binding force in virtue of helping in the pursuit of *accurate credences*

Accuracy First Epistemology

Laudable properties of credences:

- Informativeness
- Simplicity
- Unification
- Justification
- **Accuracy**

Accuracy is *the* fundamental epistemic good.

- The higher your credence in truths and the lower your credence in falsehoods, the better off you are all epistemic things considered.

Consequentialist: facts about the epistemic *good* (accuracy) explain what's epistemically *right*.

- Epistemic norms have binding force in virtue of helping in the pursuit of *accurate credences*

Epistemic Decision Theory (EpDT): Co-opt the resources of practical decision theory.

- Decision-theoretic norms explain why certain practical policies are irrational.
  - Reporting incoherent previsions
  - Economic policies, environmental policies, etc.
- Also explain why certain *epistemic* policies are irrational.
  - Having incoherent credences
  - Violating Principal Principle
  - Failing to proportion your credences to the evidence
  - Failing to update by conditionalization (except: see below)
- Bad means to the end of *accuracy*.

Epistemic rationality is a **constrained optimization problem:**

- Minimization of (estimated) inaccuracy under constraints.

Higher-Order Evidence

Higher-order evidence is evidence that you're **handling evidence in or out of accord with epistemic norms.**

Want to know how to respond to HOE in the most **accuracy-conducive way** under the **appropriate constraints.**

Hypoxia

Bob is flying his plane Tuesday morning and wonders whether **there is enough gas** to make it to Hawaii. He looks at various dials and maps, which support a **credence of .99** that there is enough gas, which Bob adopts.

Bob then receives a message from ground control that he **may have hypoxia**, which severely impairs the reliability of people's reasoning. In particular, when people with hypoxia have credence .99 in a proposition, the proposition is true only around **half the time.**

Avoiding Rationality

Will try to avoid talk of what's **rational** for Bob to do.

- Interested in question of **what Bob should do AETC** if all he cares about is accuracy.
- Want univocal answer
- 'Rational' is (possibly) equivocal and misleading
- Can still maximize estimated accuracy **under the constraint of guaranteed irrationality**.

Calibrationist All epistemic things considered, Bob should have credence less than .99.

Steadfaster All epistemic things considered, Bob should have credence .99.

Calibrationist .5 maximizes expected accuracy relative to the appropriate credence function under the appropriate constraints.

Steadfaster .99 maximizes expected accuracy relative...

Plan

Accuracy-first Defense of Calibrationism

- Claim Steadfast seems to have the upper hand *prima facie*
- Distinguish two features of HOE
- Claim troublesome feature is a kind of information loss
- Argue right constraints on optimization lead to calibrationism.

Both calibrationism and steadfastism can be thought of as means toward the end of accuracy.

- Agents who respond to HOE are more accurate than agents who don't.
- But agents who respond as Steadfasters suggest are more accurate than calibrationists.

Dispute between calibrationist and steadfaster turns on question of **estimated inaccuracy by whose lights** and **what the relevant constraints are**.

- Estimator: current cf, previous cf, ideal cf, cf matched to frequencies?
- Constraints: current evidence, actual capacity, capacities of some ideal agent, current information state?

Analogy: Newcomb's problem.

- Both CDT and EDT claim to win because they disagree about relevant comparison class.
- Steadfast and Calibrationists also win depending on how comparison is set up.

Steadfastism nonetheless seems to have the upper hand from AFE:

- Estimated inaccuracy by Bob's Monday credences.
- Estimated inaccuracy by rational urprior.
- Estimated inaccuracy by second person with Bob's evidence.

Weird Features of Calibrationism

- Failure of conditionalization
- Failure of Good's Theorem
- Apparent irrelevance
- Agent relativity
- Irrational

One Road to Steadfast

Standard norm in AFE:

ExpMin (Plan to) follow the updating procedure that minimises expected inaccuracy.

Greaves & Wallace: Conditionalization minimizes expected inaccuracy.

Greaves & Wallace

- b: The agent's (coherent) credence function
- W: Set of **epistemically possible worlds** according to b
- $\mathcal{E}$: a partition of W, s.t. the agent is sure she'll learn one proposition in $\mathcal{E}$.
- A function $r : \mathcal{E} \rightarrow$ Prob is an **updating procedure**.
- G&W show conditionalization is the updating procedure that minimizes expected inaccuracy by the lights of b.

Biased Coin?

A coin is either biased $2:1$ toward heads (B) or unbiased ($\bar{B}$).
Alice will see one flip and learn whether it lands heads H or tails T.

- $W = \{HB, H\bar{B}, TB, T\bar{B}\}$
- $\mathcal{E} = \{\{HB, H\bar{B}\}, \{TB, T\bar{B}\}\}$

Alice will end up with either $r(\{HB, H\bar{B}\})$ or $r(\{TB, T\bar{B}\})$ after learning which element of $\mathcal{E}$ is true.

$r : \mathcal{E} \rightarrow$ Prob prevents Alice from adopting different credences in the HB and $H\bar{B}$ worlds since they're elements of the same cell of the partition.

- In both the HB and the $H\bar{B}$ world, Alice ends up with the same posterior.
- **Motivation:** right constraints don't allow us to discriminate more finely between worlds than Alice herself can.
  - Can't adopt plan to have credence 1 in the true world.

- G: There's enough gas.
- E: Bob's first-order evidence.
- H: Bob has hypoxia on Tuesday.

HOE Conditionalization

In Hypoxia, conditionalization leads to ignoring HOE.

- $b_{\text{Mon}}(G|EH) = b_{\text{Mon}}(G)$

More generally, it seems Bob (before getting in the air) expects to minimize expected inaccuracy if he ignores HOE.

Higher-Order Weirdness

HOE provides both impersonal and indexical information about rationality

- The rational credence in P is at least .8.
- You might have hypoxia.

Impersonal Information

I assume some agents ought to be uncertain about what's rational.

- Assuming just one function is rational at any point, we have something like:

  RatRef $b(P|R_E(P) = x, E) = x$

- Treat rationality as an **expert**.

Nothing to See Here

The rational credence function is treated (mostly) analogously to the chance function.

- You have a distribution over possible values of R, the rational credence function.
- Gaining impersonal information about R is normal evidence.

The **indexical** component of HOE is harder to understand.

- Alice and Bob can start out with the same beliefs, learn the same things, and end up with different beliefs about impersonal facts (if Bob responds to HOE).
- This feature is responsible for the major trouble.

Quitting Certainties

Alice knows she has those weird brown beans and toast for breakfast three times a week on average.

- For any future day her credence that she has brown beans that day is 3/7.

Suppose Alice is currently certain in

- *BB*: I had brown beans for breakfast on 04 December 2016.

Alice is now planning what to think supposing she forgets *BB*.

- **Natural answer:** have credence 3/7.
- **Minimizing expected inaccuracy:** have credence 1.

This holds *ex ante* as well: What to do supposing you learn and then forget.

G&W has normative force because it gets the constraints on the optimization problem right in the appropriate cases:

- The updating plan *r* requires maintaining the same credence function in each element of a cell given partition.
- The set of possible worlds remains constant or contracts.

The problem:

- Normally: $W_{t>t_0} \subseteq W_{t_0}$.
- But in cases of forgetting, this condition does not hold.

Claim: Something similar is happening in Hypoxia.

When Bob learns he might be hypoxic on Tuesday, why doesn't he simply look at his prior on Monday and use that to decide what to do?

- $b_T(G|b_M(G|EH) = .99) = .99$
- $b_M(G|EH) = .99$

Answer: Bob on Tuesday isn't certain what his Monday prior was! So, Bob's Monday prior isn't the right one to use for expected inaccuracy minimization.

$B_{Mon} = b$: Bob's Monday prior is b.

- $W_M = \{\pm G \pm H \pm E\}$
- $W_T = \{\pm G \pm H \pm E \pm B_{Mon} = b\}$

On Tuesday, Bob has a distribution over possible values of B_{Mon} determined (at least in part) by his HOE.

More broadly, we can model some kinds of indexical HOE as (possibly) losing information about either your prior or some of your evidence

- Confirmation bias
- Survivorship bias
- Availability heuristic

Less clear if ‘calculation errors’ can fit this model, but I think so.

Basic Picture

When you gain HOE:

- The space of epistemically possibly worlds may change without contracting: $W \rightarrow W'$, but $W' \not\subseteq W$.
- Second, your distribution over possibly ideally rational credence functions changes.

What are the exact constraints? What function's estimated accuracy is to be maximized?

- When $W_t \subset W_{t_1}$, new non-extremal distribution over worlds is generated somehow.
- Just like with original distribution over W_t, no full guide on how to pick a prior over worlds.
- Treat possible past conditional credence — $B_{Mon}(\cdot|EH)$ — as expert in Hypoxia and some forgetting cases. **Reverse Reflection**
- ??? No complete story ???

What are the exact constraints? What function's estimated accuracy is to be maximized?

- When $W_t \subset W_{t_1}$, new non-extremal distribution over worlds is generated somehow.
- Just like with original distribution over W_t, no full guide on how to pick a prior over worlds.
- Treat possible past conditional credence — $B_{Mon}(\cdot|EH)$ — as expert in Hypoxia and some forgetting cases. **Reverse Reflection**
- ??? No complete story ???

Loose End

As far as Hypoxia goes, this model may work, but what about when you gain HOE about your urprior?

European Union

Bob forecasts a 20% chance of the EU dissolving by 2020 based on political data. He then learns that people with his blood type are genetically pre-disposed to being overly confident in long-lasting unions between nations. What should Bob do?

In this case, there maybe no unvarnished prior as in Hypoxia. However:

- If Bob knows his urprior and was uncertain he was rational to begin with, this evidence changes his distribution over possible rational credence functions.
- If Bob was certain he was rational to start with, then it doesn't seem like this HOE is what should sway him from that.

What Bob should do on Tuesday in Hypoxia depends on the constraints that we think are legitimate:

- The uniquely rational credence function wants Bob to ignore HOE.
- Bob-on-Monday wants Bob-on-Tuesday to ignore HOE.
- A maximally idealized agent with the same information as Bob-on-Tuesday would want Bob to respond to HOE.

But those aren't **legitimate constraints** in info-loss cases.

- Want objective function to be in same info-state.
- Unclear what right function is, but it points to some kind of calibrationism.
- Info-loss model explains oddities of HOE.