An Accuracy Argument for Self-Trust

Giacomo Molinari
giacomo.molinari@bristol.ac.uk

December 16, 2024

University of Bristol

Seld-Doubt and Self-Trust

Self-Doubt

Two kinds of self-doubt

1. **Alethic self-doubt**: doubting that my beliefs are accurate.
2. **Normative**: doubting that my beliefs are rational.

I will focus on alethic self-doubt here.

Rational Self-Doubt

It seems rational to doubt the accuracy of my own beliefs.

- **Plenty of evidence** that I have been wrong, and that my peers are wrong.
- **Preface-like cases**: I'm confident that some of my beliefs about biology are *false* (e.g. "Mammals don't lay eggs").
- **Cartesian Circle**: No non-circular way to rule out the possibility that our beliefs are thoroughly inaccurate.

Irrational Self-Doubt

Some cases of extreme self-doubt seem irrational.

E.g. believing a Moorean sentence:
"It's raining, but it's not the case that I believe it's raining".

Questions

- **Why** are certain kinds of self-doubt irrational?
- **How much** may we rationally doubt ourselves?
- What about **graded doxastic states**?
  - Being *very confident* in "It's raining, and I'm *very confident* that it's not raining" seems nearly as bad as believing a Moorean sentence.

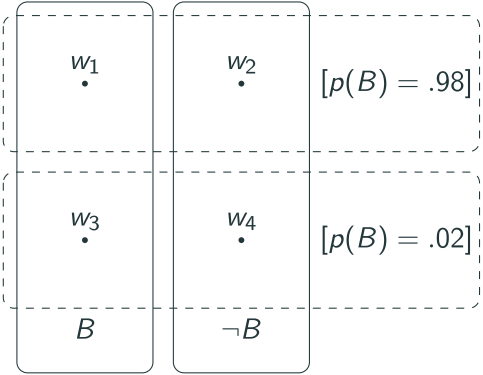
Goal: Use **accuracy** to answer these questions.

Notation

- $\mathcal{W} = \{w_1, ..., w_n\}$ finite set of *possible worlds*.
- Greek letters π, γ denote *rigidly designated credence functions*, i.e. vectors in $\mathbb{R}^n$.
- Latin letters p, q denote *definite descriptions of credence functions*.
    - p is a function from possible worlds to credence functions. So $p(w_i)$ is a credence function for every $w_i \in \mathcal{W}$.
    - Can think of them as vector-valued random variables.
    - Abuse notation: p_i instead of $p(w_i)$.
- If ϕ is a property of credence functions, $[\phi(p)]$ is the proposition $\{w_i : \phi(p_i)\}$

Example

- $p =$ My radiologist's credence function.
- $B =$ I have a broken bone.



$$p_1 = p_2 = (.97, .01, .01, .01) \;\; p_3 = p_4 = (.01, .01, .01, .97)$$

A self-trust requirement

- Let p be a definite description of your credence function.
- Let π be your actual credence function (i.e. $\pi = p_i$ where w_i is the actual world)

Total Trust

You Totally Trust yourself iff:

$$\mathbb{E}_{\pi}(X|[\mathbb{E}_p(X) \geq r]) \geq r \tag{1}$$

whenever $X : \mathcal{W} \to \mathbb{R}$, $r \in \mathbb{R}$, and the above conditional expectation is defined.

Coherence + Total Trust entails that you cannot be very highly confident of both A and $[p(A) \leq \text{low}]$.

An Accuracy Argument for Total Trust

Generalised Strictly Proper Scores

How inaccurate is expectation $\mathbb{E}_\pi(X)$ when X has value $X(w_i) = x_i$?

- Interpret $\mathbb{E}_\pi(X)$ as a *unique fair price* for gamble X.
- $(X - t)$ is desirable whenever $t < \mathbb{E}_\pi(X)$
- $(t - X)$ is desirable whenever $t > \mathbb{E}_\pi(X)$
- Inaccuracy of $\mathbb{E}_\pi(X)$ obtained by "adding up" the losses resulting from these desirability judgements.

An Accuracy Argument for Total Trust

Theorem (Dorst et al. 2012, Th.3.2)

π Totally Trusts p iff for *every* GSP measure of inaccuracy, π expects p to be at least as accurate as π.

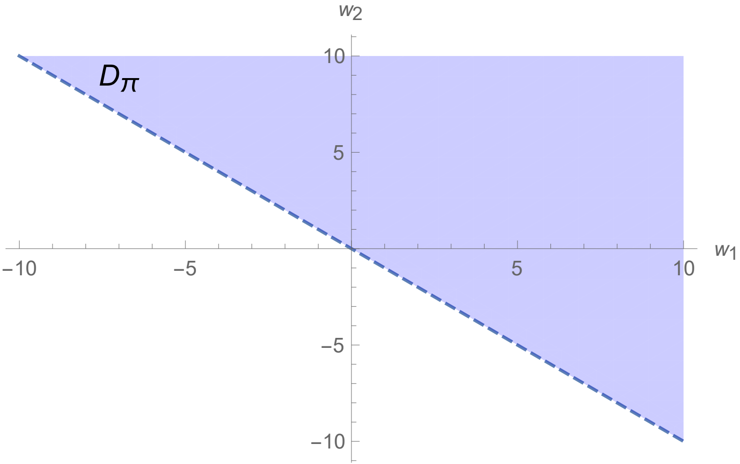
- Suppose I don't Totally Trust myself, i.e. π does not Totally Trust p.
- Then there is a rigidly designated credence function π (e.g. $(1/3, 1/3, 1/3)$) that I think is more accurate than me under some GSP measure S.
- I expect that I would be more accurate, as measured by S, by having credence function π at all possible worlds!

Problem: Why should we care about **that** measure?

Improving the Argument

From credences to desirability judgements

The desirability judgements induced by a coherent credence function π via its expectation $\mathbb{E}_\pi$, interpreted as unique fair price, are **extremely structured**.



From credences to desirability judgements

Represent opinions via **sets of desirable gambles** for more expressive power.

Subtle Point: We want to show that **rational** agents Totally Trust themselves.

- Rational agents have *coherent* and (for this talk) *precise* doxastic states.
- So we need to show that agents with *coherent*, *precise* doxastic states should trust themselves.
- Your beliefs are still representable by a coherent credence function π.
- Added expressive power lets us consider ways your beliefs **could be** that don't correspond to any coherent credence function.

From credences to desirability judgements

For any probability function π, $D_\pi = \{X : p(X) > 0\}$ is the set of gambles an agent with credence function π finds desirable.

We can express Total Trust in desirability terms.

Total Trust

π Totally Trusts p iff:

$$X \in D_{\pi(\cdot|[X \in D_p])} \tag{2}$$

whenever $X : \mathcal{W} \to \mathbb{R}$, $r \in \mathbb{R}$, and the above conditional expectation is defined.

From GSP to INSERT NAME HERE measures of inaccuracy

We can use [INSERT NAME HERE] to measure the inaccuracy of an arbitrary set of desirable gambles at a world.

- For every X which you find desirable, you get a penalty if X is not actually desirable.
- For every X which you don't find desirable, you get a penalty if X is actually desirable.

$$S(\pi, w_i) = \int_{D_{w_i} \sim D_\pi} x_i d\mu - \int_{D_\pi \sim D_{w_i}} x_i d\mu \tag{3}$$

GSP vs INSERT NAME HERE

GSP:

- **Structural Assumption**: The single value $\mathbb{E}_\pi(X)$ determines the desirability of *all gambles* in form $(X - t)$ and $(t - X)$.
- These judgements jointly determine the inaccuracy of the expectation value $\mathbb{E}_\pi(X)$.

INSERT NAME HERE

- **No structural assumptions** on desirability judgements.
- Each desirability judgement contributes *directly and individually* to your total inaccuracy.

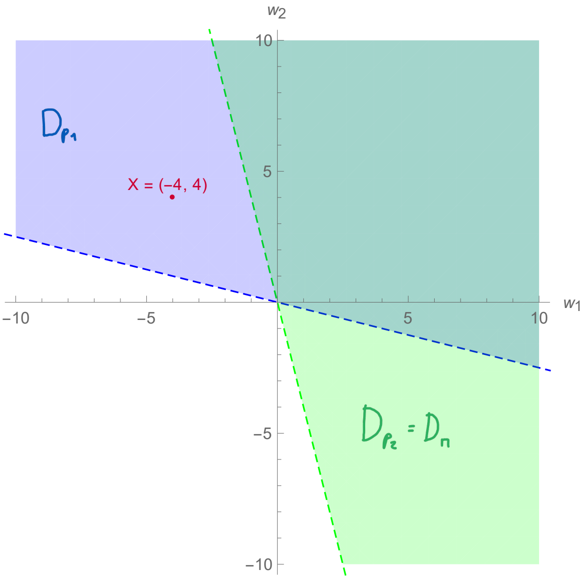
A useful fact

Fact 1

If Total Trust fails on some gamble, then it fails on some open set of gambles.

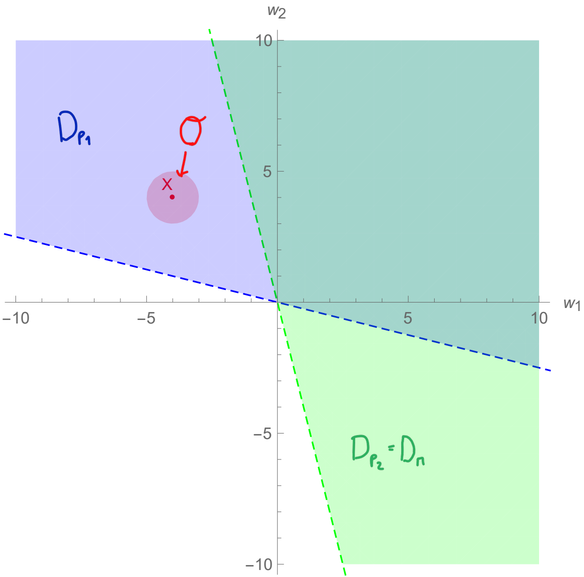
Example

- $\mathcal{W} = \{w_1, w_2\}$
- w_2 is the actual world.
- $X = (-4, 4)$



Example

- $[X \in D_p] = \{w_1\}$.
- But $X(w_1) = -4$.
- So $X \notin D_{\pi(\cdot|[X \in D_p])}$, violating Total Trust.
- Similarly for nearby gambles.



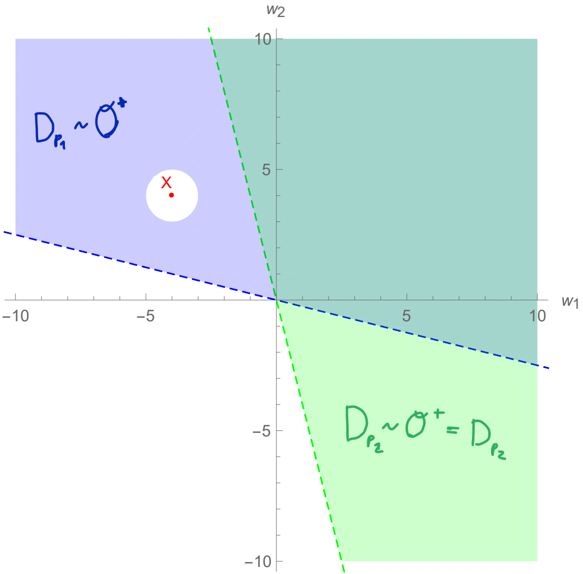
New accuracy characterisation of Total Trust

- Suppose π does not Totally Trust p.
- Then there is some open set $\mathcal{O}$ of gambles where Total Trust fails.
- Define:

$$\mathcal{O}^+ = \mathcal{O} \cap D_\pi, \quad \mathcal{O}^- = \mathcal{O} \cap D_\pi^c$$
$$D_p^* = (D_p \cup \mathcal{O}^+) \sim \mathcal{O}^-$$

- You actually find the gambles in $\mathcal{O}^+$ desirable, and those in $\mathcal{O}^-$ not desirable.
- D_p^* represents the opinions you would have if, at every possible world, you found the gambles in $\mathcal{O}^+$ desirable and those in $\mathcal{O}^-$ not desirable.

Example



New accuracy characterisation of Total Trust

- **Note**: At some possible worlds, D_p^* denotes an **incoherent** set of desirable gambles!
- But with INSERT NAME HERE we can measure its inaccuracy at all possible worlds!

Theorem

1. If π does not Totally Trust p, then there are measurable sets of gambles $\mathcal{O}^+, \mathcal{O}^-$ such that π expects D_p^* to be strictly more accurate than D_p under **every** INSERT NAME HERE measure of inaccuracy.
2. If π Totally Trusts p, then for any measurable sets of gambles $\mathcal{O}^+, \mathcal{O}^-$, π expects D_p to be at least as accurate as D_p^* under **every** INSERT NAME HERE measure of inaccuracy.

The New Argument

- Suppose π does not Totally Trust p.
- Then there are (rigidly designated!) set of gambles $\mathcal{O}^+, \mathcal{O}^-$ such that you think you would be more accurate, under **every** INSERT NAME HERE measure of inaccuracy, if you found gambles in $\mathcal{O}^+$ desirable and gambles in $\mathcal{O}^-$ not desirable at every possible world.
- There is a rigidly designated way to change your judgements which you expect would make you more accurate.

Many open questions...

- Is it really bad to expect some *possibly incoherent* definite description to be more accurate than you?
- How do we determine which doxastic states we should compare yours against when evaluating you?
- Self-trust requirements for imprecise probabilities?