An Accuracy Argument for Self-Trust

Giacomo Molinari
giacomo.molinari@bristol.ac.uk

December 9, 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 that it's raining".
$\text{Bel}(A \wedge \neg\text{Bel}(A))$

Questions

- **Why** are certain kinds of self-doubt irrational? Equivalently: why is some amount of self-trust rationally required?
- **How much** may we rationally doubt ourselves? Equivalently: what is the minimum amount of self-trust that is rationally required?
- What about **graded doxastic states**? E.g. 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.

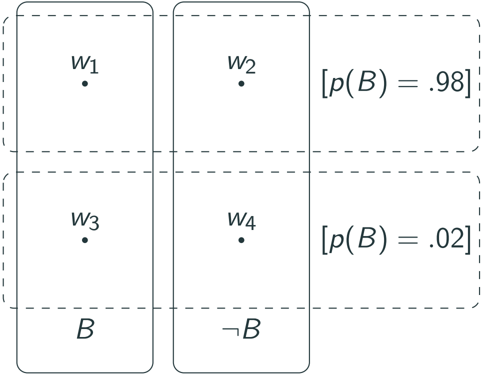
Goal: Use **accuracy** to answer these questions.

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.

I want to argue that rational agents Totally Trust themselves.

Consequences of Total Trust

Together with coherence, Total Trust entails that you cannot be maximally confident in A as well as maximally confident in $[p(A) \leq \text{low}]$, because then we would have:

$$
\begin{aligned}
&\pi(A \wedge [p(A) \leq \text{low}]) = 1 && (2)\\
\iff & \frac{\pi(A \wedge [p(A) \leq \text{low}])}{\pi([p(A) \leq \text{low}])} = 1 && (3)\\
\iff & \pi(A | [p(R) \leq \text{low}]) = 1 > \text{low} && (4)
\end{aligned}
$$

which violates Total Trust.

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

So Total Trust rules out high confidence in Moore-like sentences.

Accuracy Argument for Total Trust

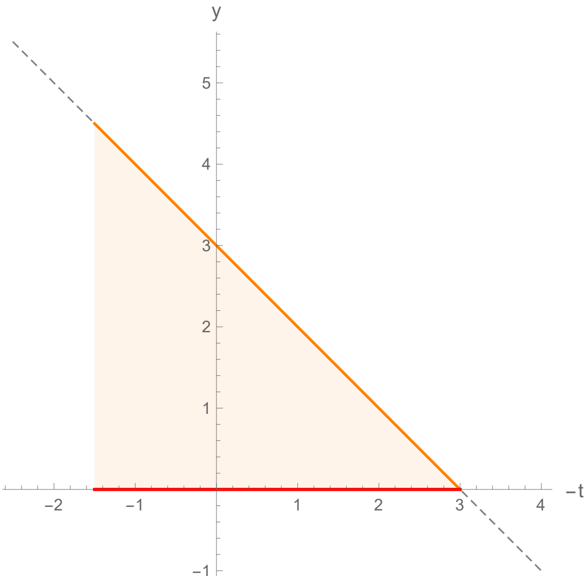
Generalised Strictly Proper Scores

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 $t < \mathbb{E}_\pi(X)$
- "Add up" the losses resulting from these desirability judgements when w_1 is the case.

Example

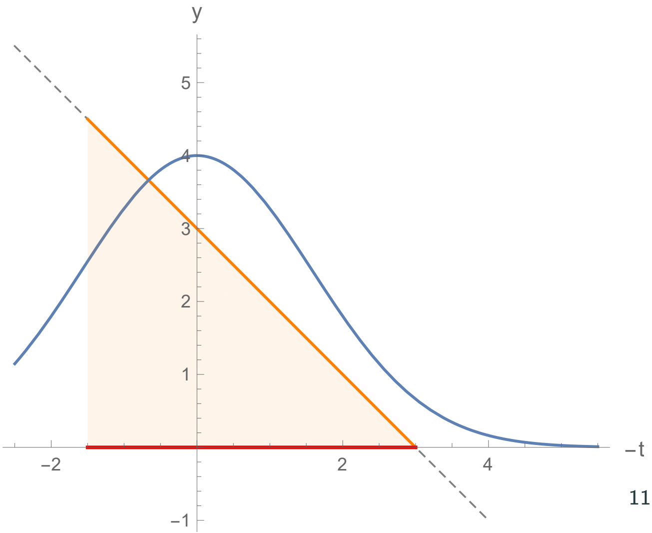
- $\mathcal{W} = \{w_1, w_2\}$.
- $X = (-3, 6)$.
- $\pi = (1/2, 1/2)$, so $\mathbb{E}_\pi(X) = 1.5$.
- Say w_1 is the case, so $X = -3$



Example

$$S(\pi(X), x_i) = \int_{x_i}^{\pi(X)} -(x_i - t)\lambda(dt)$$

Different λ yield different GSP measures of inaccuracy.



An Accuracy Argument for Total Trust

Let p be a definite description of my credence function, and π my actual credence function.

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 p under GSP measure S.
- I expect that I would be more accurate, as measured by S, by having credence function π at all possible worlds!

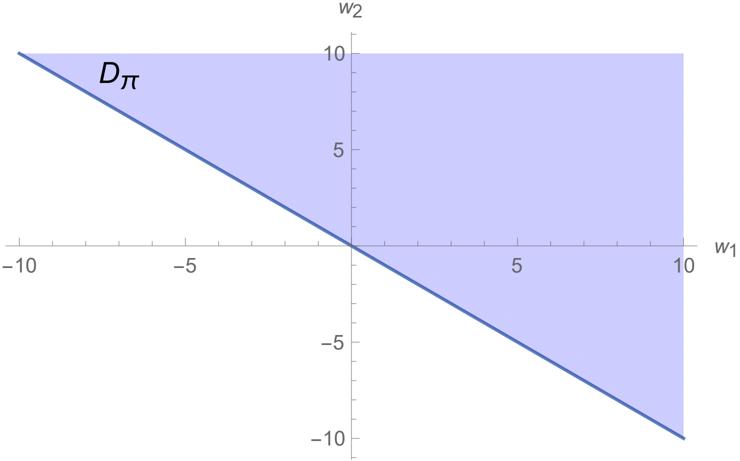
Problem

- The above theorem shows that, if I don't Totally Trust myself, then I expect some rigidly designated credence function to be more accurate than myself under **some** GSP measure of accuracy.
- But why should we care about *that* measure?

Maybe more abt why this is a problem...

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

We gain considerable expressive power by representing your attitudes via **sets of desirable gambles**.

Note: We are interested in showing that *rational* agents Totally Trust themselves.

- Rational agents have coherent and (for this talk) precise doxastic states.
- So we just need to show that agents with *coherent, precise* doxastic states should trust themselves.
- Your beliefs are still representable by a coherent credence function π.
- The 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

We can express Total Trust in desirability terms.

- Let p be a definite description of your credence function. So D_p is a definite description of a set of gambles.
- Let π be your actual credence function, rigidly designated. So D_π is a rigidly designated set of gambles.

Total Trust

You Totally Trust yourself iff:

$$X \in D_{\pi(\cdot|[X \in D_p])} \tag{5}$$

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.

$$S(\pi, w_i) = \int_{D_{w_i} \sim D_\pi} x_i d\mu - \int_{D_\pi \sim D_{w_i}} x_i d\mu \tag{6}$$

GSP vs INSERT NAME HERE

GSP:

- The single value $\mathbb{E}_\pi(X)$ determines the desirability of *all gambles* in form $(X - t)$.
- These desirability judgements determine the inaccuracy of the single estimate $\mathbb{E}_\pi(X)$.
- Structural assumption: desirable gambles are a half-space through the origin.

INSERT NAME HERE

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

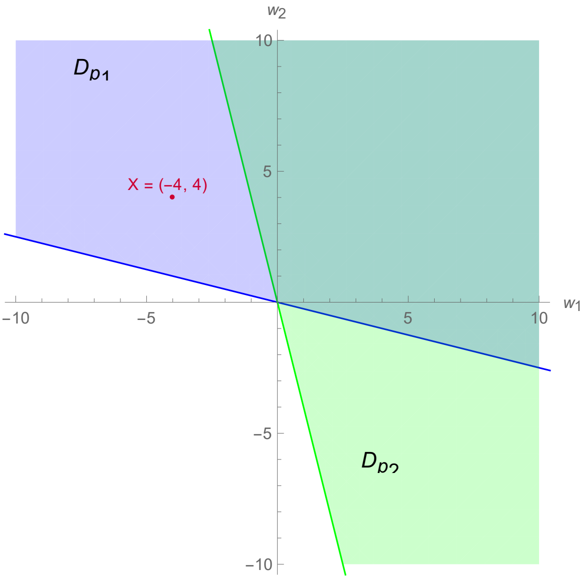
A useful fact

Fact 1

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

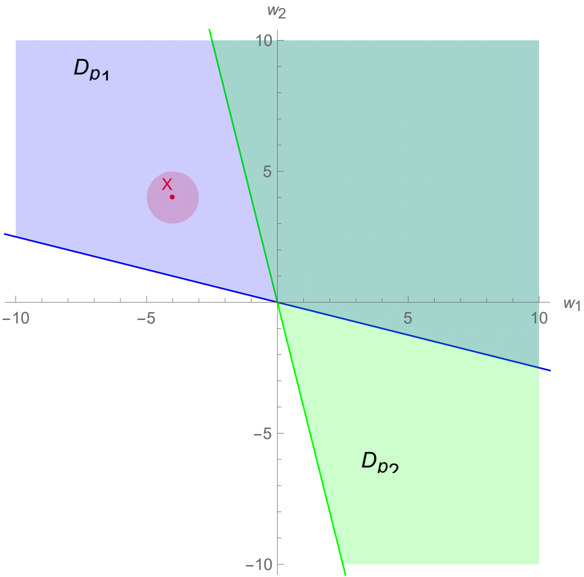
Example

- $\mathcal{W} = \{w_1, w_2\}$
- w_2 is the actual world.
- $p_1 = (.2, .8)$
- $p_2 = \pi = (.8, .2)$
- $X = (-4, 4)$



Example

- $[X \in D_p] = \{w_1\}$.
- $\pi(X|\{w_1\}) = -4$.
- So $X \notin D_{\pi(\cdot|\{w_1\})}$.
- Similarly for $X + Z$, where $-\epsilon < Z < \epsilon$



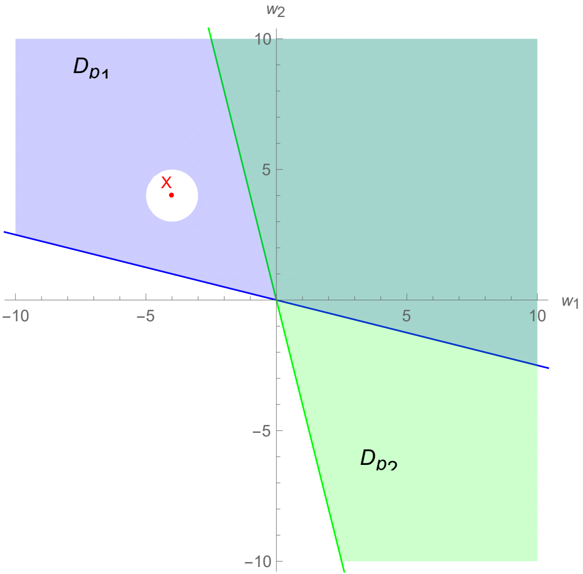
New accuracy characterisation of Total Trust

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

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

- You actually find the gambles in $\mathcal{O}^+$ desirable, and those in $\mathcal{O}^+$ not desirable.
- $D_p^{\pi\restriction\mathcal{O}}$ represents the opinions you would have if, at every possible world, you made the same desirability judgements over $\mathcal{O}$ as you actually do.

Example



New accuracy characterisation of Total Trust

- **Note**: At some possible worlds, $D_p^{\pi \upharpoonright \mathcal{O}}$ denotes an incoherent set of desirable gambles. I.e. not a half-space passing through the origin.
- 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 is some set of gambles $\mathcal{O}$ such that π expects $D_p^{\pi \upharpoonright \mathcal{O}}$ to be at least as accurate as D_p under *every* INSERT NAME HERE measure of inaccuracy.
2. If π Totally Trusts p, then there is no such set of gambles.

The New Argument

- Let *p* be a definite description of your credence function, and π your actual credence function (rigidly designated).
- Suppose π does not Totally Trust *p*.
- Then there is some (rigidly designated!) set of gambles $\mathcal{O}$ such that you think you would be more accurate, under **every** INSERT NAME HERE measure of inaccuracy, if you made the same desirability judgements as π over $\mathcal{O}$ at every possible world.
- So there is some set of gambles where you expect some rigidly designated credence function (e.g. $\pi = (1/3, 1/3, 1/3)$) to be more accurate than your own credence function *p*.
- Alternatively: you expect that, if you *determinately* made the same judgements as π over $\mathcal{O}$, you would become more accurate than you are.