Trust, Deference, and Value

Kevin Dorst
kmdorst@mit.edu

Knowledge and its Limits and its Limits
November 6, 2024

I. Rational Modesty

q = the first spoon is longer.

What credence is it rational for you to have in q? Unsure.

Let P be the rational credence function (for you, now).

> **Rational Modesty:** It can be rational to be unsure what the rational credence in q is.

$\forall t : P(P(q) = t) < 1$.

Elga's argument for modesty: hypoxia.
Williamson's: anti-luminosity.[1]

[1] Assume $Ep \Rightarrow Kp$. If $Ep \rightarrow EEp$, $Ep \rightarrow KEp$, violating anti-luminosity.

Paper's argument: disagreement.

Q: Why did past-Kevin focus on disagreement?

II. Deference: Reflection and New Reflection

Despite modesty, it's rational to (in some sense) defer to the rational credence function.

Panel of potential experts. How defer?

Let π, δ, η be particular probability functions over a finite set of worlds W. Think of them as vectors over worlds.

Eg $\delta = (0.6, 0.2, 0.2)$ over (w_1, w_2, w_3).

A set C of probability functions is *convex* iff it's closed under averaging.
Eg $\{\delta : 0.4 \leq \delta(q) \leq 0.6\}$ is convex;
while $\{\delta : \delta(q) = 0 \text{ or } \delta(q) = 1\}$ is not.

If $\delta, \eta \in C$, then $\frac{1}{2}\delta + \frac{1}{2}\eta \in C$. E.g. $\frac{1}{2}(0.6, 0.2, 0.2) + \frac{1}{2}(1, 0, 0) = (0.8, 0.1, 0.1)$.

> **Reflection:** When you learn the expert's opinion is in some range, adopt an opinion in that range.
> For any convex C: $P(\cdot | P \in C) \in C$.

Special cases:
C is $P(q) = t$
C is $l \leq P(q) \leq h$
C is $P(q) = t$ & $P(r) = t'$
C is $P = \pi$

Reflection rules out immodesty.
Example: *Immodest Imani*.

Generalizes: Elga proof

Elga's diagnosis: if you learn something about the expert, then *if they're modest*, you might've learned something they didn't know. So before deferring to them, you should *inform* them of what you've learned.

> **New Reflection:** Upon learning what the expert's opinions are, adopt the opinions they would have were they to learn what you've learned.
> $P(q | P = \pi) = \pi(q | P = \pi)$.

Define $\widehat{P}_w := P_w(\cdot | P = P_w)$ to be the *informed* rational credences.
New Reflection is equivalent to *reflecting* the *informed* credences.

For convex C, $P(\cdot | \widehat{P} \in C) \in C$

It's natural to generalize New Reflection. Let $P_r := P(\cdot | r)$. Then:

Reaction: Given q, if you're sure the (remaining) experts would react to q by adopting a given opinion, adopt that opinion.
If $P_r(l \leq P_r(q) \leq h) = 1$, then $l \leq P_r(q) \leq h$.

Letting $r := [P = \pi]$, Reaction implies New Reflection.

III. Correlation between evidence and truth

New Reflection and Reaction both allow Sycophants, hence they allow you to know the rational opinion is anti-correlated with truth.

Self-effacing evidence: $P(q \leftrightarrow P(q) < 0.5) = 1$ and $P(\neg q \leftrightarrow P(q) > 0.5) = 1$.

What's missing? Reflection implies that the rational opinions are *correlated with truth*. How do we add that to New Reflection?

In Imani case, $P(q) = 0.5$, but $P(P(q) = 0.5 \vee P(q) = 1) = 1$, so correlation implies:

$$P(q|P(q) = 0.5) < P(q|P(q) = 1)$$

$P(q)$ is an average of $P(q|P(q) = 0.5)$ and $P(q|P(q) = 1)$, so:

$$P(q|P(q) = 0.5) < P(q) < P(q|P(q) = 1)$$

But since $P(q) = 0.5$, that means $P(q|P(q) = 0.5) < P(q) = 0.5$.

Upper-bounded intervals ($l \leq P(q) \leq h$) can be evidence *against* q.

Expert might lower their credence upon learning it.

What can't? Lower-bounded intervals like $P(q) \geq t$. Points to q.

Can be formalized with vectors; $\{\delta : \delta(q) \geq t\}$ has a unique direction, and it's the same as that of $\mathbb{1}_q$.

Intuitively: upon learning that $P(q) \geq t$, you know how the expert would respond to this information—they wouldn't *lower* their credence.

So, intuitively if you learn that $P(q) \geq t$ you know (1) they started at least t, and (2) upon learning what you've learned, they wouldn't drop. That yields **Simple Trust:** $P(q|P(q) \geq t) \geq t$.

Q: Subtlety. Why doesn't Trust entail Reflection, eg $P(q|t \leq P(q) \leq t) = t$?

Formalized: minimal condition on correlation between $P(q)$ and q:

Reliance: The expert being confident of q (given r) is not evidence against q (given r).
$P_r(q|P_r(q) \geq t) \geq P_r(q)$

Reaction + Reliance is equivalent to:

Trust: If you learn that the expert reacts to r by judging that q, you should react to r by judging that q.
$P_r(q|P_r(q) \geq t) \geq t$

IV. Clockology

Trust rules out self-effacing evidence. Also rules out sure-win investigations. How?

But permits no-lose ones.

Assume (like Williamson) we model P with a *prior frame* (W, R, π).

Frame validates Trust iff $\approx$ it looks like a *tree*.

Transitive, shift-reflexive, shift-nested.

Simple creature?

Splitcon frame

Clock? Trust implies positive access.

nested best-guess model

V. Is this the right theory of deference?

Updates from 'Deference Done Better'

Trust should be **Total Trust**. Let X be a random variable (function from worlds to numbers), $\mathbb{E}_P(X)$ is the expected value of X according to P.

$\sum_w P(w) \cdot X(w)$

Total Trust: Conditional on the expert having a high estimate for X, have a high estimate for X.
$\mathbb{E}_P(X|\mathbb{E}_P(X) \geq t) \geq t$

Equivalently, for any **biconvex** set B, $P(\cdot|P \in B) \in B$.
B is biconvex if B and B^c are convex.

Features:

- Total Trust is equivalent to **Value**: no matter what decision problem you face, you'd prefer to outsource your decision to P.
- Total Trust is equivalent to *Epistemic* Value: no matter what scoring rule you use, you expect P to be more accurate than you.
- How square with Fishing for Compliments?
  $\mathbb{E}_\pi(P(q)) \neq \mathbb{E}_\pi(P^+(q))$ is consistent with Value.
  But Value forbids freely recombining access failures.

Good/bad case.

Though *question-relative* Value/Trust allows it

VI. Against Salow 2018

Salow argues that the access principle protects against IBI. Does it?

When P is higher-order certain, partitional[2] updates yield ¬IBI.

[2] I.e. access-validating

But when P is higher-order *un*certain—but satisfies access—they don't.

Know you're 50-50 on friend (f). Unsure how confident you are that *recognizable* (r)—i.e. r = *if friend, would recognize*.
You learn whether or not you recognize your friend: $\{\{rf\}, \neg\{rf\}\}$.

Note:
Prior and posterior satisfy Access.
Both satisfy Value.

$$P = \begin{pmatrix} rf & r\overline{f} & \overline{r}f & \overline{r}\overline{f} \\ 2/6 & 2/6 & 1/6 & 1/6 \\ 2/6 & 2/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 2/6 & 2/6 \\ 1/6 & 1/6 & 2/6 & 2/6 \end{pmatrix} \qquad P^+ = \begin{pmatrix} rf & r\overline{f} & \overline{r}f & \overline{r}\overline{f} \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/4 & 1/4 \\ 0 & 1/5 & 2/5 & 2/5 \\ 0 & 1/5 & 2/5 & 2/5 \end{pmatrix}$$

Suppose at *rf*. Then $E_P(P(f)) = \frac{1}{2}$. (And you know this.)
But $\mathbb{E}_P(P^+(f)) = \frac{2}{6}(1) + \frac{2}{6}(\frac{1}{4}) + \frac{1}{3}(\frac{2}{5}) = 0.55$.

Structurally:

- You're HOU about r, and we 'entangle' it with another proposition f such that you learn either $r\&f$ or $\neg(r\&f)$.
- If the former, you're certain of r.
- If the latter, that's evidence that $P(r)$ was low, which is evidence that $\neg(r\&f)$ is only *weak* evidence against f.