Deferral to experts: Combining two modelling perspectives

Leon Schöppl

May 5, 2022

Motivation

When considering the question of whether a layperson should defer to an advisor's testimony, two philosophically interesting perspectives coexist:

- The god's eye point of view with access to all relevant (and irrelevant) features of a case.
- The subjective, internal and epistemically limited perspective of the actual layperson.

I suggest employing a model that of the former perspective to *rationality benchmark* models or reasoning accounts situated in the latter.

Overview

0 *[Motivation]*
1 Duijf's model
   - Features
   - Power
   - Problems
2 Bovens' and Hartmann's model
3 The combined model
   - Objective perspective
   - Subjective perspective
   - *[Bayesian updating]*
   - What does trust mean?
   - Preliminary results
4 Research questions
5 *[References]*

Duijf's Model: Features

- A layperson L and an expert E.
- A boolean factual question φ, and a related, boolean, normative question ψ.
- L and E are correct about φ with a probability equal to their competency values l, e.
- L perfectly knows their interest whether ψ given they were correct about φ.
- E's chance of giving correct advice about ψ depends additionally on their degree of interest alignment α.

Duijf's Model: Power

The model (i) allows calculation of certain values:

- How likely is it for laypeople and experts to disagree about ψ ($p(D)$)?
  $\alpha(I(1-e)+e(1-I))+((1-\alpha)(Ie+(1-I)(1-e))$.
- What's the likelihood of incorrect expert advice ($p(I)$)?
  $\alpha(1-e)+(1-\alpha)e$.
- How likely is it that the layperson would regret deferring ($p(R)$)?
  $\alpha I(1-e)+(1-\alpha)Ie$.
- When it is rational for L to defer?
  Iff $p(L(\psi)=|\psi|) \leq p(E(\psi)=|\psi|)=\alpha e+(1-\alpha)(1-e)$.

And (ii) let's Duijf derive certain analytical results, e.g., as α increases, $P(D), P(I)$ and $P(R)$ all decrease.

Duijf's Model: Problems

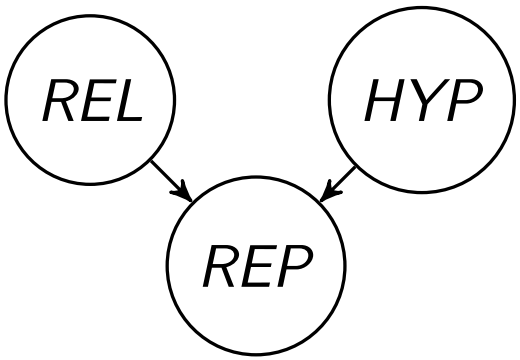
From the epistemically limited perspective, one cannot access the values required for these calculations, undermining (i).

In addition, much of the analytically derived results are contingent on overly idealized assumptions or on an inaccessible perspective, undermining (ii):

- When regarding a specific domain, a layperson's competence may well be worse than chance.
- From the perspective of actual laypeople, an advisor may not be an expert at all, but a *charlatan* with $e < l \vee e < 0.5$.
- Not only do laypeople lack (perfect) access to e, but also to α.

As a result, the model is not feasible as a heuristic employed by laypeople, and most of the analytically derivable results break down.

Bovens' and Hartmann's model



- Reliability: $p(REL) = r$
- Hypothesis: $p(HYP) = h$
- Report:
  - $p(REP|HYP, REL) = 1$
  - $p(REP|\neg HYP, REL) = 0$
  - $p(REP|HYP, \neg REL) = \beta$
  - $p(REL|\neg HYP, \neg REL) = \beta$

The combined model: Objective perspective

- Pairs of experts and laypeople, with the values needed for Duijf's model (e, l, α).
- One factual proposition φ, and one normative question ψ (the value of the latter depends on each L).
- Laypeople and experts assess φ (based on l, e) and experts give advice on ψ based on α and their assessment of φ.
- Using Duijf's formula, the model determines whether deferral is objectively rational in each case.

The combined model: Subjective perspective

Laypeople employ the Bayesian model, with the following initializations:

- $p(HYP|\psi) = I, p(HYP|\neg\psi) = 1 - I$ (So I set $h = I$ or $h = 1 - I$).
- Based their *astuteness*, laypeople take a guess as to e, α and calculate the chance of correct expert advice (cr) using Duijf's model.
- Then $p(REL) = |(cr - 0.5) \times 2|$, with either (i) inverting the testimony of presumed worse-than-chance advisors, or (ii) setting their reliability to 0.
- $p(REP)$ is determined as usual, with $\beta = 0.5$ in the case of (i) or based on cr and HYP in case (ii).

Bayesian updating

(1) $p(REP) = 1rh + 0r(1-h) + (1-r)h\beta + (1-r)(1-h)\beta = rh + \beta - r\beta$ (Law of total probability)

(2) $p(REP|HYP) = r + (1-r)\beta = r + \beta - r\beta$ (Law of total probability)

(3) $p(HYP|REP) = \frac{p(REP|Hyp)p(HYP)}{p(REP)} = \frac{(r+\beta-r\beta)h}{rh+\beta-r\beta}$ (from 1 and 2, Bayes rule)

(4) $p(HYP|\neg REP) = \frac{p(\neg REP|Hyp)p(HYP)}{p(\neg REP)} = \frac{(1-(r+\beta-r\beta))h}{1-(rh+\beta-r\beta)}$ (from 3, Negation rule)

(5) $p(REP|REL) = 1h + 0(1-h) = h$ (Law of total probability)

(6) $p(REL|REP) = \frac{p(REP|REL)p(REL)}{p(REP)} = \frac{hr}{rh+\beta-r\beta}$ (from 1 and 5, Bayes rule)

(7) $p(REL|\neg REP) = \frac{p(\neg REP|REL)p(REL)}{p(\neg REP)} = \frac{(1-h)r}{1-(rh+\beta-r\beta)}$ (from 6, Negation rule)

The combined model: What does trust mean?

Three conditions for interpreting a layperson's behavior as *trusting* their advisor:

- L adjusts their belief in HYP in the direction of E's advice.
- L ends up with a $p(HYP) \geq 0.5$ given the expert gave testimony that ψ (and $p(HYP) \leq 0.5$ otherwise).
- L increases their $p(REL)$ following their updating on E's testimony.

Preliminary results

I've run *Behavior space* for the settings most faithful to Duijf's paper, for a total of 100000 layperson-expert pairs:

- Both laypeople and their advisors are better than chance at determining whether $\varphi : e, l \in (0.5, 1]$.
- The advisors are more competent than the laypeople: $e > l$.
- The degree of interest alignment ranges from none to perfect: $\alpha \in [0, 1]$.

It turns out that for these settings, the Bayesian laypeople acted rational according to Duijf's model in a mean 72.4% of cases, in just the first round.

Research questions

- Implement repeated updating, including *regret*.
- Explore how robustly these Bayesian agents perform well for different value ranges of l, e, α and *astuteness*.
- *Rationality benchmarking* additional ways of Bayesian updating on (lack of) testimony, e.g. à la Hartmann & Heinzelmann.

References

- Luc Bovens and Stephan Hartmann (2004) Bayesian epistemology.
- Hein Duijf (2021) Should one trust experts?
- My model: `https://github.com/leon-schoeppl/abm-experts`
- Stephan Hartmann and Nora Heinzelmann (2022) Deliberation and Confidence Change.