DEFERRAL TO EXPERTS: COMBINING TWO MODELLING PERSPECTIVES

Leon Schöppl

March 2022, Draft

INTRODUCTION

Broadly, there are two different angles from which to approach the question of whether an agent should defer to the testimony of another: On the one hand, what I will call the *god's eye perspective*, making use of all the facts about the situation to draw a definitive, objective conclusion. On the other hand, the subjective, internal and epistemically limited perspective of the actual agent in question, the inhabitants of which need to make use of evidence and a variety of heuristics to determine how to proceed.

In this short essay, I want to argue that a specific model of layperson-expert deferral, namely the one introduced in Duijf 2021, should be clearly situated in the god's eye perspective, as it appears unusable as a heuristic from within the internal perspective[1]. Nevertheless, I will argue, the model can be fruitfully employed by formal epistemologists, if in a rather niche fashion, namely to model a testing environment for other, actually subjective models. I created a proof of concept by writing just such a model, in which agents deliberate trust in each other using (subjective) Bayesian updating à la Bovens and Hartmann 2004, and their success (or failure) is determined using objective assessments à la Duijf. While the scope of this project doesn't allow a deep-dive into the model's predictions, one preliminary result, that this combination of two modelling perspectives enables me to draw, points to Bayesian laypeople being quite promising doxastic agents for this specific epistemic problem.

I will first briefly contextualize the target phenomenon of all three models featured in this essay — laypeople-expert trust deliberation — in the social-epistemology literature. Next, I present an overview of the elements and technical workings of Duijf's model[2], giving an argument why it should be considered to explicate the god's eye perspective in the process. I will go on to introduce the model for deferral to the testimony of a witness by Bovens and Hartmann, attempting to make precise, how it differs in per-

[1] This assessment builds on worries already raised by Duijf in that very paper.

[2] Using my own notation in the process.

spective from the Duijf's. Finally, I will introduce the features of my model, explaining how it squares the differences between both of its predecessors, before concluding with one preliminary result and a possible further research question.

Background

Social epistemology can be defined as the "enterprise concerned with how people can best pursue the truth [...] with the help of, or in the face of, others." (see Goldman and O'Connor 2021), and in deliberating whether to defer to the testimony of another, this aptly frames the choice one has to make: Is whoever is offering advice a helpful source of knowledge, or rather incompetent and-or deceptive to the point where trusting them would hinder one's inquiry?

In deliberating these questions, social epistemology sometimes takes a perspective that is quite foreign to traditional epistemology, treating collectives or systems as the agents and central structures of the *social* within which inquiry occurs. In the way it is modelled by Duijf though, the phenomenon of layperson-expert deferral belongs to that variety of social epistemology concerned with *individual doxastic agents* (see A. Goldman 2010), so that my model too will centrally feature individual experts and laypeople, who are merely engaged in communication with one another.

As these (individual) agents are by no means all knowing, there are two related questions about the rationality of trusting someone's testimony that need disentangling: On the one hand, one might be interested in whether — given all the facts about a situation — a layperson should trust an expert. To adequately be able to answer this, one is required to know not only how competent the layperson and experts are in answering factual questions, but also whether they are honest and communicate effectively. In other words, one needs to inhabit the god's eye perspective. On the other hand, one might be interested in the fallible, internal perspective of the layperson considering trusting their vis-à-vis, and all the epistemic difficulties that come with that. No longer blessed with unhindered access to the truth of the questions deliberated, nor to the competencies of either of the people involved, the central issue now becomes whether the layperson has sufficient evidence and justification to put trust in the proclaimed expert.

I would judge the second perspective to be more central, after all it is the one that we all constantly inhabit. Hence, it is unsurprising that there exists a wide-spread social-epistemology literature on how laypeople are justified in trusting despite the risks involved (see e.g., Lackey 2010), which heuristics they may employ to distinguish between more or less trustworthy experts (see e.g., A. I. Goldman 2001), and how to handle testimony from people one suspects to be trustworthy only to limited degrees (see e.g., Elga 2007). As should become clear from my description of it in the next section though, the model introduced in Duijf 2021 appears to rather be an exception: As it requires quite generous access to facts that are generally inaccessible from the internal perspective, it should be regarded as modelling the god's eye perspective instead.

Duijf's Model

Duijf 2021 introduces a model of expert deferral containing a designated layperson L and a designated expert E, each with their own competency values (l, e). When deliberating a central, boolean and factual question φ, the expert and layperson are correct with probability e, l respectively. Assuming that both experts and laypeople are better than chance at such deliberations, and that experts are better than laypeople, Duijf restricts these competency values s.t. $l, e \in (0.5, 1]$ and $e > l$.

Following the assessment of φ, the layperson now faces a choice on a related, normative question ψ, where the value of ψ ($|\psi|$) is relative to the interests of L. Assuming that laypeople are in tune with their own interests, the probability that L makes a correct ψ-choice ($L(\psi) = |\psi|$) is 1 given they were correct on the underlying question whether φ ($L(\varphi) = |\varphi|$), and 0 otherwise. For the expert, though, their own interest may align with that of the layperson to a higher or lesser degree ($\alpha \in [0, 1]$), meaning that the probability that the expert gives correct — relative to the laypersons interests — advice on ψ is α given the expert deliberated correctly about φ, and $1 - \alpha$ otherwise.

In other words, even if the expert and layperson agree about the facts (φ), the expert may suggest a different path of action (ψ), if their interests do not match those of the layperson. At the same time, the two might also come to different conclusions about φ but agree as to what the layperson ought to do about ψ. This dynamic becomes especially interesting, as the expert is not presumed to communicate their assessment of φ to the layperson, instead merely giving advice on ψ.

The model allows Duijf to derive probabilistic results about a variety of scenarios:

- How likely is it that the expert and layperson disagree as to what choice the layperson should make ($p(E(\psi) \neq L(\psi))$)? This comes down to $\alpha(l(1-e) + e(1-l)) + ((1-\alpha)(le + (1-l)(1-e))$, and can come about in two different ways: One the one hand, the expert and layperson could already disagree about φ and their interest is sufficiently aligned that this disagreement carries over to the normative question ψ, which is reflected by $\alpha(l(1-e) + e(1-l))$, on the other hand they might agree about φ, but their misaligned interests turn that unity into a disagreement about ψ, explicated by $((1-\alpha)(le + (1-l)(1-e))$.

- With what probability does the expert give incorrect advice ($p(E(\psi) \neq |\psi|)$)? This comes down to $\alpha(1-e) + (1-\alpha)e$, namely the probability that the expert is incorrect about φ but sports a sufficiently misaligned interest, plus the probability that they are already incorrect about φ, and hence commit what one might call an honest mistake. Correct expert advice ($p(E(\psi) = |\psi|)$) on the other hand, is simply $1 - p(E(\psi) \neq |\psi|)$.

- With what likelihood would the scenario merit the layperson regretting having deferred to the expert's testimony $(p((L(\psi) = |\psi|), (E(\psi) \neq |\psi|))$? This is determined by $\alpha l(1-e) + (1-\alpha)le$, namely the probability that the layperson is correct, while the expert is incorrect about φ and hence gives incorrect advice about ψ ($\alpha l(1-e)$), plus the probability that both the layperson and expert are correct about φ, but the expert gives incorrect advice due to misaligned interest ($(1-\alpha)le$).

- When, or given which values of l, e, α, is it advisable for a layperson to trust an expert? Iff $p(L(\psi) = |\psi|) > p(E(\psi) = |\psi|) = \alpha e + (1-\alpha)(1-e)$, that is when the layperson is more likely to be correct about what to do than the expert to give correct advice.

Despite the many simplifying assumptions that go into this model, I take it to be a powerful tool at determining under what conditions a layperson ought to defer to an expert's testimony. And yet, there are some fundamental problems in the way of simply applying it from the perspective of an actual layperson:

In addition to the above, Duijf (using partial derivatives) derives some further results about the dynamics of the model: As the expert's degree of interest alignment with the layperson increases, the chance of incorrect expert advice, disagreement and regret go down. Yet, these results are contingent on the (relative) restrictions placed on the competency values of both experts and laypeople[3], which we ought to reconsider from the internal perspective of the layperson.

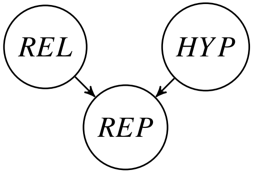
Laypeople are not necessarily blessed with a competency value that is better than chance, at least not for any possible factual question. Even if, on average, they are better than chance on factual questions throughout their lives, as soon as they consider a φ from an unfamiliar domain, that is no longer the case. And it appears that determining one's own competency level for any specific question φ would be — if not outright impossible — then at least similarly error-prone as determining the answer to φ in the first place. Combine that with the problem of determining whether the person proclaiming to be an expert actually is one, and not an impostor, and the whole situation is blown wide open. From the perspective of the layperson, then, almost everything goes: Both people involved in the exchange may have competency values in the full unit interval.

Add to this Duijf's own results that it is impossible for laypeople to determine the exact values of α and e from an expert's testimony w.r.t. ψ alone, and practically unfeasible to determine α even given e, and the use case of this model as a heuristic for laypeople can be ruled out.

[3] As I've already shown in my midterm essay.

Boven’s and Hartmann’s model

The Bayesian Model (see p. 69) of Bovens and Hartmann 2004 on the other hand feels entirely at home in the layperson’s perspective. It is meant to explicate how one can investigate a hypothesis using testimony from a (more or less) reliable witness. In terms of Duijf’s target phenomenon, I will understand it as a model of how a layperson can form and update their beliefs on a question ψ based on the testimony of an alleged expert. Represented by the following Bayes Net[4], it features three nodes:



- Reliability (*REL*), a root node with probability r, representing the likelyhood that the (presumed) expert is a reliable witness, thereby combining both e and α from Duijf’s model into one subjective probability assigned by the layperson.[5]

- In the way I want to employ the model, the root-node Hypothesis (*HYP*) simply represents ψ, or more precisely, the proposition that ψ (rather than $\neg\psi$) is in the layperson’s interest ($|\psi| = true$). The prior probability of HYP is simply given by a fixed prior probability h in the original model.

- Report (*REP*), representing the presence of a witness report as to the truth of HYP, or in my interpretation, expert testimony that ψ is in the laypersons interest. Being a child node of both REL and HYP, the probability of REP is given in the original model by the following distribution (where β is a randomization parameter):

  - $p(REP|HYP,REL) = 1$
  - $p(REP|\neg HYP,REL) = 0$
  - $p(REP|HYP,\neg REL) = \beta$
  - $p(REL|\neg HYP,\neg REL) = \beta$

  As might become clear from these conditional probabilities, reliability is understood differently in this model than competency is in Duijf’s[6], namely as the question whether testimony of a witness can be epistemically useful at all (see p. 57) (even as higher order evidence). An unreliable witness as understood in this way is not consistently wrong, but instead just gives randomized testimony, reporting with probability β as to the truth of HYP.

[4] Some helpful basic familiarity with the technical notions may be gained in Bovens and Hartmann 2004, p. 67.

[5] At this point, there admittedly exists some friction between the two models, which I will discuss in the subsection *The internal perspective*.

[6] Another point of friction which will be addressed in the subsection *The internal perspective*.

Using Bayesian updating, this model allows Bovens and Hartmann to adequately spell out how a rational agent (or Bayesian layperson) should reconsider their beliefs in how reliable their vis-à-vis is and how likely ψ is, based on whether the alleged expert gives testimony that ψ or that $\neg\psi$.

The combined model

Implemented in *Netlogo* (Wilensky 1999) *6.1.1*, my agent-based model (Schöppl 2022) aims to combine the perspective of Duijf's model with that of Bovens and Hartmann, or in other words, let the two interplay with each other. The model features the (objective) values required for Duijf's model to assess the rationality of layperson-expert deferral on the one hand, and Bayesian laypeople with the subjective credences required by the model of Bovens and Hartmann on the other. In the following two subsections, I will give a quick overview of the details (and potential problems) of my implementation.

The god's eye perspective

The model features a boolean proposition φ with an objective value, and an adjustable (in the user interface *UI*) number of linked expert-laypeople pairs. Laypeople and experts each are initialized with their respective competency values (l, e) which remain unchanged throughout a single run of the model. Experts feature a fixed degree of interest alignment α with their layperson. The possible values for l, e, α can be adjusted in the UI, where I also implemented a toggle for the assumption that (for each expert-layperson-pair individually) $e > l$. For each individual layperson, there is an objective truth about their interest as to ψ (given $\varphi, \neg\varphi$ respectively), and the correctness of an expert's advice will be measured against these individual interests.

Based on these features and values, experts and laypeople each assess the truth of φ, which they are correct about with a probability equal to their competencies. In addition, experts will give testimony as to whether ψ, based on α and their assessment of φ. Using Duijf's model, I also determine (for each layperson-expert-pair), whether it would be rational for the layperson to trust the expert. These results are reflected in the UI, where agents that correctly assess φ are represented by Smileys (by Xs otherwise), experts that give correct advice to their laypeople are coloured green (red otherwise) and links that between laypeople and trustworthy experts are coloured green (red otherwise).

The internal perspective

In addition to these objective facts about the model and its inhabitants, laypeople also feature a few subjective credences, as well as the mechanisms to reason from them in a Bayesian fashion. To that end, I've implemented the nodes from Hartmann's and Bovens' model, and initialized them as follows:

Realistically, as noted above, a layperson should not have direct access to the truth of φ, nor to their own competence l as to figuring that out. Still, $p(HYP)$, which appears as a root-node with a simple prior probability to the layperson, can nevertheless be defined in terms of these objective values in the combined model: $p(HYP|\psi) = l, p(HYP|\neg\psi) = 1 - l$. To see why this works, recall that a layperson figures out the truth about φ with probability l, and — as they have perfect interest alignment with themselves — also figure out the(ir normative) truth about ψ with that same probability. Hence, I initialized the HYP node for each layperson individually by setting $h = l$ (or $h = (1 - l)$).

In addition, there initially exists a mismatch between the expert reliability in the Bayesian Model, and the probability of correct expert advice in Duijf's model. To square the two, I implemented the following solution[7]: I initialized laypeople with prior beliefs in e and α, and depending on a *layperson-astuteness*-slider in the UI, these beliefs can range from highly accurate (about the respective expert), to entirely up to chance.[8] Astuteness hereby defines an interval centred on the objective values of e, α within which the layperson's guess lays. E.g., for an astuteness value of 0.2 and $e = 0.5$, the laypersons guess as to e is $\in [0.3, 0.7]$

Based on e and α, the layperson starts out with an initial belief in the expert giving correct advice ($cr(E(\psi) = |\psi|)$, abbreviated here as cr), which I calculated using Duijf's results. Then, the layperson's prior in REL is determined by $|(cr - 0.5) * 2|$. Thus, e.g., an expert with a 50% chance of giving correct advice is reliable with probability 0, one with 75% chance of giving correct advice with probability 0.5 and one with 100% chance with probability 1. Given the symmetry of this initialization, I simply set the randomization parameter β to 0.5.

Note here that laypeople will treat experts with a below 50% chance of giving correct advice as reliable higher order evidence, that is, both recognizing them as *Charlatans*, but also making use of the fact that their inverted testimony can be taken to be probably correct. While I chose this way of implementing laypeople to accurately reflect Bovens and Hartmann understanding of what it means for a witness to be reliable, I do recognize that it does not match well with Duijf's understanding of rational deferral: If an expert's interest is sufficiently misaligned that they give incorrect advice to their layperson, then even if L may update on this testimony, they should surely not *defer* to it. Therefore, I have implemented an option to *not update on reliably bad experts*, in which anyone suspected to be such a specimen earns a $p(REL) = 0$. In this case, β is initialized with the likelihood with which the layperson would expect their vis-à-vis to advise them to do ψ, based on their own assessment of whether φ.

Laypeople use Bayesian updating to re-evaluate their credences in the reliability of their expert (REL) and truth of ψ (HYP). They arrive at posterior credences in REL and HYP depending on whether

[7] Admittedly, this is somewhat sketchy.
[8] One might imagine more astute laypeople to be employing heuristics such as the ones proposed in A. I. Goldman 2001.

they receive testimony of the expert as to do ψ (*REP*) or as to not do ψ ($\neg$*REP*). For how exactly the updating procedures work, see the *appendix*.

In this (early version of the) combined model, I offer a two part understanding of a layperson to be trusting their vis-à-vis[9], which is reflected in the UI by colouring the relevant layperson green (red otherwise). The first part requires a layperson to (i) adjust their belief in HYP in the direction of their expert's advice and (ii) end up with a $p(HYP) \geq 0.5$ given the expert gave testimony that ψ (and $p(HYP) \leq 0.5$ otherwise). The second part requires L increasing their belief in *REL* following their updating on *E*'s testimony.

With this compromise between the two previous models, it is now possible to model and observe cases of laypeople-expert deferral for a variety of input values, and judge the choices and behaviour of the Bayesian laypeople according to the objective predictions made by Duijf's model. For each run, the model reports how many interactions between laypeople and experts have occurred, and in how many of those the laypeople came to act rationally in terms of Duijf's model. The next subsection will give a very brief overview of one resulting observation.

A (preliminary) result

As mentioned in the introduction, a full exploration of the combined model's predictions is not feasible here. Nevertheless, I want to briefly point to one result that I find particularly interesting:

Using *behavior space*, I have run 1000 simulations — containing 10 expert-layperson pairs each — of settings that I deem as faithful as possible to Duijf 2021, using the limits to the parameter intervals that his original model expects: $e, l \in (0.5, 1], \alpha \in [0, 1], e > l$. That is, both experts and laypeople are better than chance, and an expert is always better than their respective layperson, and the degree of interest alignment can range from 0 all the way up to 1. In choosing a *layperson astuteness* value of 0.2, I attempted to strike a balance between the worries that laypeople cannot avoid falling for charlatans on the one hand, and the optimism arising from the plethora of secondary resources of trust that A. I. Goldman 2001 points to: While a layperson can never be absolutely certain how competent their vis-à-vis is, they can investigate their reputation amongst other (meta-)experts, track-record, biases. Given that updating on the inverted testimony of what one figures to be reliably bad experts is hardly trust, and doesn't match Duijf's interpretation of deferral to (the actual) expert testimony, I have instead selected the option in which such experts are treated as fully unreliable.

Out of the 10 deliberations in each run, in a mean of 7.495 whether the layperson trusted their expert concurred with what Duijf's model deemed rational from the god's eye perspective. That is, these Bayesian laypeople with their limited access to the objective facts of the situation were objectively

[9] The UI allows selection not only of this combined notion of trust, but also of each individual version, by changing the variable *What does trust mean?*.

rational in 75% of cases![10] As such, given rather conservative assumptions about laypeople's astuteness in the face of proclaimed experts, my model predicts Bayesian laypeople to be considerably better than chance at determining whether to trust in their vis-à-vis[11].

Conclusion and outlook

Despite it being hardly employable as a heuristic by actual laypeople, there appears to exist a fruitful *way out* for Hein Duijf's model of layperson-expert deferral. Unlike the epistemic predicaments individual doxastic agents constantly face in social situations, employing it to evaluate the reasoning strategies of Bayesian laypeople allows the model to do what it does best: Computing epistemically usually inaccessible values about competencies and interest alignment, and describing objectively whether it is rational for a layperson to defer to an expert on a specific matter.

For the specific parameter settings most faithful to Duijf, my combined model's Bayesian laypeople fared quite well at trusting those experts that Duijf's model deems trustworthy.

Given the limited scope of a term-paper project, my model is still in its infancy. The next step in refining and extending it would be to implement a way for laypeople to update on whether they regret their decisions of (dis-)trusting their vis-à-vis, by allowing them to retroactively gather evidence about (or simply learn) whether φ, and thus evaluate the testimony they received against a more reliable data set than their own predictions.

From there on, I am optimistic that the combined model can be used to derive (further) new and interesting results about the phenomenon that neither Duijf's nor Hartmann's and Bovens' models could deliver on their own.

References

Bovens, Luc, and Stephan Hartmann. 2004. *Bayesian epistemology.* OUP Oxford.

Duijf, Hein. 2021. "Should one trust experts?" *Synthese* 199 (3): 9289–9312.

Elga, Adam. 2007. "Reflection and disagreement." *Noûs* 41 (3): 478–502.

Goldman, Alvin. 2010. "Systems-oriented social epistemology." *Oxford studies in epistemology* 3:189–214.

Goldman, Alvin, and Cailin O'Connor. 2021. "Social Epistemology." In *The Stanford Encyclopedia of Philosophy,* Winter 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

[10] Contingent on my combining of these two models properly encoding what it means to trust an expert.

[11] For these specific parameter choices.

Goldman, Alvin I. 2001. "Experts: Which ones should you trust?" *Philosophy and phenomenological research* 63 (1): 85–110.

Lackey, Jennifer. 2010. "Testimony: Acquiring knowledge from others."

Schöppl, Leon. 2022. "ABM-experts v1.0," https://github.com/leon-schoeppl/abm-experts/releases/tag/v1.0.

Wilensky, U. 1999. "NetLogo," https://ccl.northwestern.edu/netlogo/.

Appendix

(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)