Master's Thesis

Interpreting and evaluating Bayesian source-reliability models

Munich Center for Mathematical Philosophy
Ludwig-Maximilians-Universität München

Klee Schöppl

Munich, February 20th, 2023

Submitted in partial fulfilment of the requirements for the degree of M. A.
Supervised by Prof. Dr. Stephan Hartmann and Dr. Hein Duijf

This page was intentionally left blank.

Contents

1 Introduction

Imagine—as a children's riddle might start—that you stand before two unmarked stone doors and are forced to make a critical choice, as behind one of them awaits good fortune, and behind the other certain death. Imagine further that the doors are guarded by one of a variety of gargoyles: They may be a truth-telling knight or lying knave, and they may know exactly which door is which or be utterly confounded as to that fact. Now, you might not know which kind of gargoyle you are facing, and hence, how good their advice will be. But you can be sure that it will be either of the following four cases: if the doors are guarded by a well-informed knight, then you are in luck: Competent and truth-telling as they are, you can follow their advice as to which door to enter through. Confounded knights, while well-meaning, are nevertheless anti-reliable advisors, guaranteed to send you on the wrong path. Similarly, well-informed knaves will be anti-reliable qua attempting to deceive you, and you would be well advised not to trust them. Finally, what to make of the confounded knave? They are utterly wrong about the matter at hand, but as they are simultaneously attempting to deceive, their advice will be just as helpful as that of the well-informed knight: Their double anti-reliability cancels itself out.

Of course, testimonial situations are rarely so black-and-white: advisors need neither be perfect truth-tellers, nor utterly committed to lying, and few matters are so simple that competent advisors get them correct 100% of the time, or so difficult and confusing that incompetent advisors never stumble upon the right answer. Right in between these extremes, for example, we might imagine a new variety of gargoyle which simply bases their advice on a coin flip they made before your arrival at the stone doors. Such randomizing gargoyles would be just as unreliable as you already are on your own, and following their advice would maintain your original $50 : 50$ chance of opening the correct door. From the subjective perspective, as the recipient of advice, the task at hand is as clear as it is difficult: assess the competence of an advisor and figure out, for instance, whether they truly are the expert they pretend to be or merely a charlatan. And even once that question is answered, you are left to determine whether your advisor's testimony may be tainted by deception.

Bayesian models of source-reliability are formal descriptions meant to capture this

perspective: based on the subjective probabilities (credences) that a recipient assigns to the relevant parameters of the testimonial situation, they determine how these credences ought to be updated upon receiving evidence in the form of testimony. Over the last two decades, these normative models have been used to explain, evaluate or rationally reconstruct various reasoning-related behaviours and phenomena. For example, they have been applied to observed responses to conjunction fallacy tasks, and used to philosophize about peer disagreement and belief divergence (see Bovens and Hartmann 2004, Olsson 2020a, Heinzelmann and Hartmann 2022, Henderson and Gebharter 2021). They have also been implemented as the mechanism to model trust in larger scale network simulations to study macro-phenomena like the formation of filter bubbles, polarization or the formation of correct or incorrect consensuses (Olsson 2011, Olsson 2013, Angere 2010, Assaad 2022, Hahn, Hansen, and Olsson 2020, Olsson 2020b), and have received some empirical support from the literature on argumentation and on the psychology of reasoning (Hahn and Oaksford 2007, Hahn, Harris, and Corner 2009, Oaksford and Hahn 2012, A. J. Harris et al. 2016).

To these various ends, these models are often not applied as stand-alone tools, but rather either extended to form the core of more extensive Bayesian models or implemented as one of the multiple gears in larger computer simulations. In both cases, accurately understanding these comparatively simple models is a fundamental first step to correctly assessing the behaviour of the complex systems that employ them, and to ensuring that we draw correct conclusions from the results they produce. Given the range of applications and the potential impact they could have, we ought to be exceptionally thorough in our discussion of these models.

This thesis continues this project, which two recent papers (Hahn, Merdes, and Sydow 2018, Merdes, Sydow, and Hahn 2021) have started: it offers a unified way of interpreting and modelling the testimonial situation formally, introduces and compares various relevant source-reliability models based on this perspective, and evaluates their behaviours and limits both conceptually, and with the help of computational simulations.

In the second chapter, I present a non-Bayesian model of the testimonial situation based on work by Hein Duijf, defining source-reliability as a function of an advisor's competency and degree of interest alignment with the recipient. After pinning down the details of how to properly interpret this model, I go on to discuss the limits of its application: drawing from Alvin Goldman's work on 'veritism', I situate Duijf's in a 'God's-eye' perspective on the testimonial situation, and warrant the need for other models to inhabit the subjective perspective of the recipient.

As Bayesian source-reliability models are the perfect candidate for such a shift in perspective, chapter three starts out with an introduction (to the technical details) of Bayesian reasoning with a particular focus on two aspects: expectation based updating on testimony received from an advisor and belief revision for fine-grained second-order credences. From there I take the time to introduce (variations of) four simple Bayesian source-reliability models, explaining and contrasting them using the results from the previous chapter: the original model introduced by Bovens and Hartmann in 2004; the model introduced by Olsson and Angere as part of the popular network simulation software *Laputa*, which features fine-grained reliability estimations and recognizes the possibility of anti-reliable sources; a variation independently introduced by Assaad and A. J. Harris et al. to disentangle source-competency from source-alignment; and finally, my own model created to combine the best aspects of its predecessors into a 'Bayesian mirror' of the objective perspective described by Duijf's model.

The fourth chapter serves to conceptually clarify a variety of aspects of these focal source-reliability models, starting with an explanation of how to translate the competency and alignment values from Duijf's model—and back. Next comes the process of excavating a cogent notion of 'trustworthiness-accuracy' to evaluate these models, during which many of their individual shortcomings are highlighted. These differences and shortcomings should lead us, as I argue in section 4.3, to select the overall most apt of these source-reliability models, lest we run the risk of ad hoc hand-crafting models fine-tuned to fit specific application contexts. Section 4.4 further disentangles the difference between applying these models globally, i.e. as part of a larger and more complex Bayesian network, or locally, well suited to the application in computer simulations, before Section 4.5 raises two problems specific to works following the latter approach.

And finally, in chapter five I use computer simulations to visualize and explore many of the features, flaws and behaviours I explained and detailed throughout the thesis: how do the order effects incurred by local model applications arise? Are there ways of bench-marking these models despite the limitations of expectation-based source-reliability estimation? What role can the 'randomization parameter' play, a feature specific to two of these focal models? How do these models differ with respect to the steadfastness of their source-reliability estimation, and why?

2 Modelling the testimonial situation

In this chapter, with the help of a model developed by Duijf, I introduce a way to represent the testimonial situation formally, based on the advisor's competency, the

recipient's competency, and on the degree to which both their interest align. Next follows a more detailed interpretation of this representation, linking its formal components to features of the target phenomenon. In the third section, drawing on work from Goldman, I discuss the limits this kind of model incurs when applied to actual instances of deliberation of trust in advisors.

2.1 Hein Duijf's objective model

Duijf 2021 introduces a simple model of the testimonial situation, requiring only three values ($\in [0,1]$) for its representation thereof: the competency of the advisor κ, the competency of the recipient ρ, and degree of interest-alignment α between the two parties. While the advisor and recipient each figure out the truth about the matter at hand with probabilities equal to κ, ρ respectively, the advisor's testimony is additionally influenced by α to be correct with probability equal to $\alpha\kappa + \overline{\alpha\kappa}$.[1] [2]

Duijf takes the testimonial situation to feature two related, Boolean propositions, a factual φ and a contingent, normative ψ. Applied to the simple introductory example, φ would be the answer to the question *Which door is which?*, and ψ answers which door you should open. Having perfect interest alignment with yourself, upon deciding your beliefs about φ, you will immediately derive a belief about ψ. The gargoyles, however, may—if they are knaves—flip their factual assessment of φ due to their interest alignment of $\alpha = 0$.

Notice that as soon as either κ, α go to 0.5, the probability of correct advice ($\alpha\kappa + \overline{\alpha\kappa}$) does as well, and as a result, the advisor is categorized as a randomizer. This is intuitively correct: if an advisor flips a coin when determining whether φ, their alignment with the recipient becomes irrelevant, the randomness on the level of φ will pervade the later reasoning about ψ. And, conversely, no matter how good an advisor is at deliberating φ, if they then throw a coin to determine whether they will accurately report these results or invert them, this later randomness overrides their competency-based inquiry. See Figure 1 for a detailed overview of how exactly the values of κ, α relate to types of advisors. Duijf provides some additional insight into the dynamics of this model using partial derivatives: $\dfrac{\partial p(C)}{\partial \alpha} = 2\kappa - 1$ tells us that the probability of correct testimony increases with increasing α iff the advisor is more competent than

1. $\overline{x}$ is short for $(1 - x)$ throughout this thesis.

2. I want to quickly mention that Duijf also provides formulae to calculate two further values, which will not be as relevant for this thesis: the probability of disagreement between advisor and recipient $\alpha(\rho\overline{\kappa} + \kappa\overline{\rho}) + \overline{\alpha}(\rho\kappa + \overline{\rho\kappa})$, and of the recipient being in a position to regret deferral to their advisors, that is, the probability that the recipient is correct about ψ, while the advisor is incorrect: $\alpha\rho\overline{\kappa} + \overline{\alpha}\rho\kappa$.

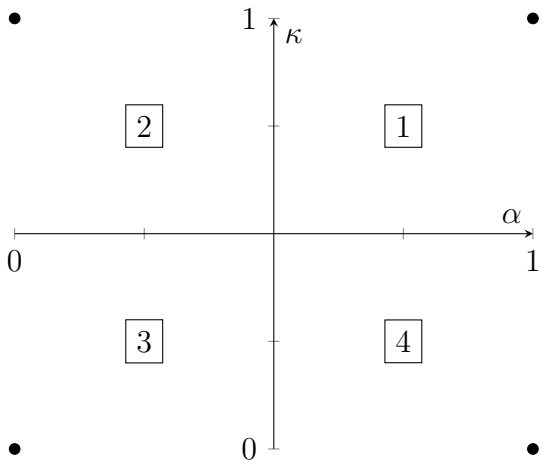

Figure 1: Full possible spectrum of advisors, based on their competency κ and alignment α. Advisors on the horizontal axis are competency-based randomizers, and those on the vertical axis randomize based on alignment. Advisors in the first quadrant are more reliable than chance, qua having better than chance competency and above 0.5 alignment. Advisors in the second or fourth quadrant are anti-reliable due to their below 0.5 alignment and worse-than-chance competency, respectively. The third quadrant represents advisors that are more reliable than chance qua being double-anti-reliable. The dots mark the extreme cases, which are perfectly reliable sources for quadrants one and three, and perfectly anti-reliable sources for quadrants two and four.

chance. Similarly, vice versa, the probability of correct testimony increases together with increasing κ iff the interests align to more than 50% ($\frac{\partial p(C)}{\partial \kappa} = 2\alpha - 1$).[3]

As introduced in Duijf 2021, the values for κ, ρ were originally restricted to be within $(0.5, 1]$, taking both advisor and recipient to always be better than chance in their individual inquiry. As such, the original model only captures the first two quadrants and avoids dealing with double anti-reliability altogether. Duijf even put the additional restriction of $\kappa > \rho$ in place, as the model was meant to be applied only to experts advising laypeople, resulting in a reasonable assumption that the advisor is more competent than the recipient. Contingently on this combination of restrictions, Duijf was able to calculate further analytic results about the model's behaviour, e.g., how changing α impacts the probability of disagreement between recipient and advisor (see Appendix-section 7.1 for more details).

I do, however, take below-chance competency to be a real possibility: advisors with this unfortunate feature are a natural result of testimonial chains which allow for testimony that is anti-reliable due to anti-alignment (e.g., lying). For example, assume

3. Duijf actually shows the inverse of this behaviour for the probability of incorrect testimony $p(I)$, but these reformulations follow directly from the fact that the probability of correct testimony is $1 - p(I)$.

your advisor has gathered all their information about a domain of inquiry from a single source, like a newspaper, then there are at least two ways their resulting competency may be below chance: (i) they may simply trust, and update on, the information provided by a competent, but anti-aligned source, and (ii) they may distrust, and hence anti-update on, the information provided by a competent, aligned source. But apart from deliberate misinformation or defective testimonial chains, below-chance competency may also be the result of harbouring incorrect background assumptions: for an extreme example, assume an advisor is of the belief that the 'nature versus nurture' debate about the relative causal impact of genes and environment on the development and success of human beings was to be settled 100% in favour of nature. Such an advisor might then take any occurrence of societal inequalities as direct evidence for genetic differences, rather than as evidence for changeable environmental inequalities. As a result, it would be unsurprising if they exhibited competency-based anti-reliability with respect to questions about social and economic policies aiming at ameliorating inequalities. And, as Figure 1 shows, whenever below-chance κ is combined with anti-alignment like when, e.g., the above, news-paper reading advisor tries to deceive their recipient, the respective advisor becomes more reliable than chance.

Apart from providing conceptual clarity about the testimonial situation, the biggest advantage of Duijf's model is that it allows calculation of whether a recipient should defer to the testimony of an advisor: this, according to Duijf, is the case iff the chance of correct advice is greater than the recipient's own competency (iff $\alpha\kappa + \overline{\alpha\kappa} > \rho$). Similarly, advice may be fruitfully incorporated in a recipient's deliberation process iff it is at least better than chance ($\alpha\kappa + \overline{\alpha\kappa} > 0.5$).

For example, assume a recipient has no clue about the answer to a Boolean question φ, and is aware of that fact: the best they can do on their own is to flip a coin ($\rho = 0.5$). Even if imperfect both in competency and in their degree of interest alignment, an advisor may still turn out to be more reliable than this recipient. For example, e.g., $\kappa = 0.8, \alpha = 0.7$ give rise to a $56,6\%$ probability of correct advice. In this specific example, the recipient would be wise to defer to the received advice. And even if ρ were above $56,6\%$, there may still be ways for the recipient to aggregate their own assessment and the received testimony in epistemically fruitful ways.

2.2 Model interpretation

So far, all of the central elements of the testimonial situation have remained rather unspecified in the *OD*-model, namely the natures of the *question(s) at hand*, of the

competency values κ, ρ, and of *interest alignment*. This section resolves this by offering a more detailed interpretation of each of them.

Let us start with defining φ, ψ as Boolean propositions, either completely true or completely false. We can then use them to model yes/no-questions of the sort "Is it the case that φ?" (Goldman 1999, p. 91), and issues as to whether φ or $\neg\varphi$ is the case (Goldman 2001b, p. 242). In answering the question of how to understand competency, I suggest we follow Goldman's characterization of expertise—a special case of competency—as a sort of "veritistic propensity" (1999, p. 91). He defines expertise as an objective notion, as a measure of how many propositions of a relevant domain a person holds true (or false) beliefs about, which also aligns perfectly with A. J. Harris et al. 2016, where expertise is understood and modelled as a field-specific property of advisors. In 2018 (p. 5), Goldman defines that someone is "an expert about [a] domain D if and only if (A) [they have] more true beliefs (or high credences) in propositions concerning D than most people do, and fewer false beliefs; and (B) the absolute number of true beliefs [they have] about propositions in D is very substantial."

(A) here is a relative notion, which makes sense for expertise specifically, but can be dropped when defining a simple degree of competency: I will understand the competency values κ and ρ as relative to a domain of inquiry Φ, and as expressing the objective chance of an advisor or recipient correctly assessing the value of an arbitrary $\varphi \in \Phi$. φ might be a new question previously not considered by an advisor, perhaps a forecast or prediction, or even just as a randomly chosen question $\in D$ which they have already made up their mind about. This definition avoids misconceptualizing competency in an overly specific manner, e.g., as pertaining only to a single question at hand, which would make the application of the model fruitless and render competency equivalent to simply 'being correct'. Understanding competencies in this way also allows for people to vary in their competency between different areas of inquiry: for example, if presented with an arbitrary question in their area of expertise, a highly competent advisor may well have the disposition to get it right in 95% of cases, meriting a $\kappa = 0.95$, while simultaneously having lower competency values associated with different areas of competence.

A quick note on the constitution of such domains of inquiry: assume φ_n is one of the n tokens of the type of questions under consideration in a domain Φ, then 50% of the elements of Φ are true and when we pick an arbitrary $\varphi \in \Phi$, it has a 50% chance of being true (and false) respectively. This is because each Boolean question in any domain can simply be negated to arrive at a related Boolean question with an inverted truth value, meaning that for any given domain of questions, I take it that any randomizing strategy, like (weighted) coin flips, will always converge to 50% accuracy

for questions in one domain. This does not however mean that they necessarily result in the same ratio between false-positive testimony and false-negative testimony.

Duijf's distinction between φ and ψ as factual and normative faces a variety of problems: firstly, it burdens us with making a clear distinction between the two spheres, which the literature on the fact-value dichotomy proves to be difficult (see e.g., Väyrynen 2021). Secondly, even if we were able to clearly make that distinction, it is doubtful that testimony given by advisors exclusively concerns normative questions, meaning the model could not be applied to the entirety of the testimonial situation. Thirdly, even assuming a clear fact-value dichotomy and that all testimony is about value questions, it is unclear why every Boolean normative issue should be grounded in exactly one factual question so that the truth value of ψ depends solely on that of φ. Fourthly, this interpretation would have to leave room for an additional variable representing the degree of interest alignment a layperson has with themselves, so that e.g., even when the layperson is correct about φ, it is not necessarily the case that $|\psi| = L(\psi)$.[4] And lastly, let us quickly consider the application of the model to multiple recipients $R_1 ... R_n$ of testimony, which becomes relevant when attempting to understand more complex social situations: should ψ be understood as the general proposition that some choice, course of action, or policy is good simpliciter, with the value of ψ depending on the actual interest of specific recipients? In this case, there could be disagreements of the form $R_1(\psi) \neq R_2(\psi)$, and φ too would have to be a quite general factual question, as to the entirety of impacts of that policy. Alternatively, ψ might be understood as agent-relative, as a collection of specific propositions of the form *some policy is good for some agent*, writing distinct interests with respect to the policy as $|\psi_{R_1}| \neq |\psi_{R_2}|$. Lastly, there is the option of taking ψ_{L_n} to be the piece of advice that lays in the interest of a specific recipient given the value of φ is settled. This would simplify modelling insofar as that the value of ψ is the same for each agent in the model but comes at a cost too: the semantic content of two propositions $\psi_{L_n}, \psi_{L_m} \in \Psi$ are now distinct in an opaque way far and beyond just being the interest of different agents with respect

4. For example, why not allow for a recipient who learns the impact of an economic policy on different tax brackets to commit an error in estimating whether the policy is in their own interest, e.g., by being too optimistic about which tax bracket they may find themselves in, a few years down the line? One might try to square this by arguing that if the normative question is truly contingent on the factual question about how the world is, then any details of how it impacts the layperson would have to be part of the factual assessment φ. In this example, this would mean that if ψ is the question of whether a certain tax policy is good for a working-class recipient, then that recipient's competency value ρ would have to encapsulate not merely their likelihood of determining the benefits of that policy for different classes, but should also already include their propensity to inaccurately consider themselves upper-middle-class, or even a future millionaire. However, this would be an ad hoc inflation of the semantic contents of φ.

to a single policy.

An easier and better way of interpreting the model, then, is to understand ψ simply as an abbreviation for the public deliberation of φ, so that $A(\varphi)$ is the advisor's factual assessment of the question, while $A(\psi)$ is their public testimony about it. Conversely, $R(\varphi)$ would be the recipient's factual assessment of the question, with $R(\psi)$ (or simply $|\psi|$) being their interest as to what kind of testimony they want to receive about it. In this interpretation of ψ as purely epistemic, the recipient's interest always coincides with the value of φ. Here we simply understand them to want accurate testimony about whether the correct answer to a Boolean question is 'yes' or 'no', and the two-step modelling via φ, ψ serves to make explicit how the expert arrives at (i) their assessment of and (ii) their testimony about this question. Re-applied to the introductory gargoyle example, this interpretation of the OD-model already proves more fruitful: previously, we were only able to model a gargoyle giving advice on which door you should go through, whereas now, we may also represent the question as to which door is which. For either question, the gargoyle's competency κ determines their internal assessment of the question (φ), and their degree of interest alignment α determines their testimony (ψ).

This epistemic interpretation of φ, ψ also simplifies the notion of a degree of interest alignment significantly. Where previously there was room for advice to diverge from the advisor's assessment of φ due to misaligned background assumptions, the degree of interest alignment between an advisor and recipient now simply captures notions of sincerity, honesty or trustworthiness, which is again in line with Goldman (see 1999, p.123).[5] Of course, when seeking to apply this model, one faces the empirical task of identifying suitable domains of inquiry, be they scientific disciplines, questions answered using a specific type of methodology or posed with the aim of a specific project. However, providing a detailed account of how to properly classify domains of inquiry is beyond the scope of this thesis. Just as a consideration: it might make sense to further divide a domain of inquiry into subdomains for which it is possible to determine uniform degrees of interest alignment. Take the example of a general practitioner (GP) advising their patient on matters of health. At first glance, one might model any advice given by the GP as pertaining to questions in the same domain. However, it might well be that the doctor's interest alignment is not so simple:

5. Goldman 1999 (p. 123) suggests further disentangling competency from opportunity, whereas in my understanding, an advisor without the opportunity to receive or reflect information about a domain is simply less competent with respect to it.

In matters of life and death, their alignment will be perfect, equal to 1. However, in relatively low-stakes questions, e.g., when prescribing one of two brands of similarly effective headache medication, the GP's financial interest may outweigh their patient's interest more often than not. As a result, it may be most fruitful to model low-stakes and high-stakes medical questions using different values of interest alignment α.

2.3 Veritism and perspective problems

In his 1999 book *Knowledge in a social world*, Goldman introduces and defends veritism, the idea that because "[p]eople have interests, both intrinsic and extrinsic, in acquiring knowledge (true belief) and avoiding error" we should evaluate "intellectual practices by their causal contribution to knowledge and error" (p. 69), and "select the social practices that would best advance the cause of knowledge" (p. 79). More specifically, he proposes we should measure a person's knowledge and error with respect to a question at hand using veritistic value, and evaluate epistemic practices by the expected change in veritistic value they produce (see p. 89).

As mentioned, Goldman, too, is concerned with Boolean questions or issues of the form *whether* φ *or* $\neg\varphi$. Agents contemplating such issues entertain a subjective degree of belief (credence) p in both the truth of φ and in the truth of $\neg\varphi$ (hopefully equal to $1-p$). Being Boolean, however, φ is either objectively true or false, and the verisimilitude of these credences depends on this truth value. Let us say that $|\varphi| = 1$ if φ is true, and $|\varphi| = 0$ otherwise. Then the veritistic value of a credence p in φ is $|\varphi|p + \overline{|\varphi|}\overline{p}$. For example, when first encountering the gargoyles, you had a degree of belief $p = 0.7$ that you should step through the left of the two doors. If that is indeed the case, your veritistic value with respect to this issue is 0.7, and it is 0.3 if you should actually enter the door on the right instead.

According to Goldman, "[p]ractices have *instrumental* veritistic value insofar as they promote or impede the acquisition of fundamental veritistic value" (p. 87), and he distinguishes between *target practices*, amongst which veritism tries to select the best ones, and *selection practices*, which are used to choose among them (p. 79). Simplified examples of the former include lying, speaking the truth and believing the testimony of advisors. The latter, among others, might include the creation of and philosophizing about source-reliability models.

Recall that in Section 2.1, I already argued against the possibility of restricting the values of κ, ρ to > 0.5 a priori. Additionally, there is also a good reason to relax Duijf's second restriction of $\kappa > \rho$: from the subjective perspective of the recipient, the

pre-selection of expert advisors is a task just as challenging as sorting out misaligned advisors. Goldman (2018, p. 6) calls this the expert-identification problem: properly defining competency and expertise is one thing, but identifying the genuine experts—as opposed to charlatans—is another. He has spent much time thinking and writing about this issue: in Goldman 2001a (p. 93 ff.), he identifies various avenues for recipients to evaluate the competency and interest alignment of advisors. Alongside the primary source for trust, the actual arguments presented by advisors to support their position, come secondary ones, like an advisor's track record, academic standing and evidence about their biases. In 1999 he categorizes these sources of trust as remedies against scepticism either in a reporter's competency, or in their honesty, but then goes on to say that "[u]nfortunately, many of these efforts to enhance credibility can be duplicated or approximated by deceptive and even incompetent reporters. Inevitably, the task of deciding how much confidence to place in a report falls to each hearer" (p. 108).

Upon relaxing both restrictions, Duijf's analytic results are no longer straightforwardly applicable (see again Appendix-section 7.1). This makes the model considerably less helpful from the subjective perspective of a recipient: not knowing the exact competency and alignment values of your advisor and being uncertain about your own competency values, you cannot use the OD-model to determine the chance of correct expert advice, nor whether you should defer to this advice over your own assessment. Being further unable to restrict the possible values to $\kappa > \rho > 0.5$ you cannot use the model's analytic results to, for example, know a priori how the chance of correct expert advice develops as interest alignment increases.[6]

This is an instance of a general problem also faced by Goldman: "[t]he implementation of veritistic epistemology is difficult for still another obvious reason. In defining the [veritistic values] of belief states and (derivatively) of practices, I assumed that the beliefs have objective truth values. This assumption does not imply, however, that those truth-values are *known* to the veritistic theorist, or that they are easy to ascertain." (1999, p. 91). What is needed, is a change of perspective: "Facts may simply be illusive; especially when they pertain to non-observable matters. But this itself does not undercut the appropriateness of veritistic criteria, especially when viewed from a "God's eye" perspective" (2018, p. 5). Where from the perspective of a recipient of testimony seeking to veritistically evaluate their listening practices his theory mostly provides "*conceptual* clarity" (1999, 91), things are vastly different from this objective,

6. In adjusting the model to be able to represent the full spectrum of conceptually possible advisors, I am not committing to the view that in reality, advisors should be expected to be uniformly distributed across it. Instead, I suspect such distributions to heavily depend on the types of questions deliberated, and the stakes involved.

God's eye perspective, whence veritism allows categorization and selection of listening practices. And, in the same manner, the God's eye perspective provides a home to the *OD*-model, and allows its application to determine cases of rational deferral.

Olsson (2011) combines these ideas already, by suggesting that computational modelling can step into the intersection between these two perspectives to play an epistemically fruitful mediator role: as modellers, we have direct access to the 'God's eye' perspective in the little worlds which we create, and thus access to objective chances and truth values that are obscured from the subjective perspective of applicants of an epistemic practice, which Olsson calls the *determination problem*. When it comes to veritism, however, he argues that computational exploration can tackle a second issue, the *computational problem*, and cites Goldman:

> "Veritistic social epistemology seeks to assess not only the practices currently employed by people and communities but to inquire whether there might be better practices to replace those presently in use. This means that practices must be evaluated that, so far, have no track record at all. To evaluate such hitherto undeployed practices, one must consider how they would perform in a range of possible applications. In other words, we must consider their veritistic "propensities," not just their veritistic "frequencies." In fact, the same point holds for practices that do have a prior track record. Whatever that track record is, it may be partly due to various accidental features, which are not firm guides to the future performance of the practice. Needless to say, it is not easy to determine the prospective performance of a practice. It cannot be determined by direct empirical observation, only by theoretical considerations, typically conjoined with background empirical information. This makes the task of veritistic epistemology extremely difficult." (1999, p. 91)

Once one has implemented a suitable epistemic environment to automatically calculate (changes in) veritistic values of the beliefs held by focal, simulated agents, the computational problem can be tackled head-on: automatically running simulations for much of, if not the entire parameter space of the model can be handled with minimal human effort and input. Varied variables might include the sizes and compositions of the simulated epistemic communities, the epistemic practices they employ and the truth values of the questions investigated. Many of the papers discussed in this thesis follow Olsson lead, some even specifically using or referencing *Laputa*, the software created by him and Angere (Vallinder and Olsson 2014, Pallavicini, Hallsson, and Kappel 2021, Olsson 2020b, Hahn, Hansen, and Olsson 2020, Assaad 2022), or the source-reliability model contained therein (Hahn, Merdes, and Sydow 2018, Merdes, Sydow, and Hahn 2021). Evaluating the veritistic values across many simulations for the same parameter choices requires a scoring rule to account for the inherent stochasticity of

each individual result. For example, an advisor with $\alpha = 0.8, \kappa = 0.8$ assesses Boolean questions correctly with a probability of 0.68, but for any ten consecutive assessments, there is much variety in how many pieces of each type of advice are given, and in which order. The measure of choice for many of these papers is the so-called 'Brier score' (Brier et al. 1950), or mean squared difference between the objective value of φ ($|\varphi| \in \{0, 1\}$) and the agent's credence c in φ for all n agents: $\frac{1}{n}\sum_{i=1}^{n}(|\varphi| - c_i)^2$. In terms of Goldman, this is just the mean squared veritistic value.

3 Focal Bayesian models

In this section, we switch the perspective from the 'God's eye' view on the testimonial situation that the objective model afforded us, to the perspective inhabited by the recipients of advice. I will start out by introducing the concept of Bayesian agents and Bayesian rationality, considering some of their limitations in the process. Next, I will go on to introduce Boolean and continuous variations of four different focal Bayesian models of source-reliability, which all differ from each other with respect to how much of the model space described by the *OD*-model they can capture and represent. The chapter concludes with a brief overview of focal models that will not be centrally featured for the remainder of the thesis.

3.1 Bayesian rationality

As we have just discussed, when receiving advice, from a gargoyle for instance, one inhabits a perspective that does not lend itself to the application of veritism nor the *OD*-model. Luckily, Bayesian reasoning is perfectly at home in this subjective perspective. In this section, I will very briefly introduce the notions of Bayesian conditionalization, Bayesian networks and expectation-based updating, which are necessary for understanding Bayesian source-reliability models.

Bayesian agents assign subjective degrees of belief ('credences') in propositions: "a person's credence in X is a measure of the extent to which she is disposed to presuppose X in her theoretical and practical reasoning" (Joyce 2005, p. 154). The baseline of rational Bayesian reasoning[7] is the fulfilment of two conditions: the credences entertained (i) must satisfy axioms of the probability calculus, and (ii) they must be updated using Bayesian conditionalization. That is, every event is assigned a subjective probability value in the unit interval $[0, 1]$, the total amount of probability adds up to 1, and for two

7. For an in-depth explication of the Bayesian approach to reasoning, see Howson and Urbach 2006.

mutually inconsistent events A, B (so that $A \wedge B \equiv \bot$), $p(A \vee B) = p(A) + p(B)$. Belief revision is then handled using Bayes' formula: $p(H) = \frac{p(E|H)p(H)}{p(E)}$, with $p(E)$ being given by $p(E|H)p(H) + p(E|\neg H)p(\neg H)$ for Boolean hypotheses. Upon encountering a new piece of evidence, Bayesian agents update their priors in related propositions based on the conditional probability that E occurs given that H is true, divided by E's unconditional probability.

In a nutshell, there are two general approaches to arguing for Bayesian reasoning: on the one hand, arguments from practical rationality warn us that reasoners who disobey the two conditions run the risk of being 'Dutch-booked', of entering into a set of bets that guarantees a profit for their bookie, and a loss for themselves. Epistemic arguments, on the other hand, promise that if "an agent's subjective likelihoods exactly match the objective likelihoods, then conditionalizing on [the evidence] will have an (objectively) expected increase in truth possession" (2001, p. 243), even if there can obviously be no guarantee that of increases in all cases (Goldman 1999, p. 123 ff.). The caveat in this promise lies with the condition of knowing the objective likelihoods, though Goldman later argues that "comparatively accurate subjective likelihoods (or subjective likelihood ratios in the same direction from 1 as the corresponding objective ratios) still promote the truth-acquisition process" (p. 251). He draws the following analogy:

> "Valid deductive reasoning, by itself, does not help a reasoner get to truth (true conclusions). Unless reasoning starts with *true premises*, deductive reasoning offers no guarantee, or even hint, that conclusions will be true. Basically the same holds for Bayesian reasoning. Simply conforming subjective probabilities to the rules of the probability calculus offers no guarantee that conclusions will be true, or even probably true. An extra something must be satisfied in the Bayesian reasoning process, if it is going to have any propensity to move one towards true conclusions. The extra something I propose is that the agent's likelihoods should be at least approximately accurate." (Ibid., p 240)

These likelihoods, and in particular the likelihood ratio $\frac{p(E|H)}{p(E|\neg H)}$, essentially describe the diagnosticity of the evidence (Hahn, Harris, and Corner 2009, Hahn and Oaksford 2007), and specifically, when approaching the testimonial situation from a Bayesian perspective, the diagnosticity of testimony, which in turn depends on the reliability of our advisor.[8] Goldman (1999, p. 113) explicates this in the following formulation of Bayesian conditionalization on confirmatory testimony:

8. For more details about the Bayesian perspective on source-reliability in the context of argumentation, see Hahn, Oaksford, and Harris 2013, Oaksford and Hahn 2012. The first of these two chapters offers a more general introduction, whereas the second contains an experimental study of how source-reliability-based reasoning interacts with the acceptance of ad hominem arguments.

$$p(X) = \frac{p(Testimony(X)|X) \times p(X)}{p(Testimony(X)|X) \times p(X) + p(Testimony(X)|\neg X) \times p(\neg X)}$$

For example, as you happen upon the two unmarked stone doors from the introduction, entirely unable to determine which is which, you might as well flip a coin. However, given your certain background knowledge that it must be exactly either of these doors, your credence that it is the one you want to enter will be exactly 0.5. Let us assume that you are advised by a gargoyle that you know to be a perfectly reliable, well-informed knight, and they tell you to enter the left door. In this case, the conditional probability that this gargoyle would advise you to enter the left door given this is actually the cogent choice is exactly one, and the probability that they would do so otherwise is zero. As a result, updating in the way Goldman suggests leads you to a posterior credence in 'I should enter the left door' equal to 1. Exactly as one would expect, the procedure suggests deferral to a perfectly reliable source. However, given that the problem with source-reliability is that it tends to be obscured from the subjective perspective, how do you arrive at these conditional probabilities and their likelihood ratio?

Because Bayesian conditionalization requires the use of accurate likelihoods to be epistemically fruitful, we need a systematic way of accurately estimating the diagnosticity of testimony in relation to our current beliefs about the reliability of our advisor, and the hypothesis in question. We can do this by specifying the probabilistic dependencies between (1) the truth about the issue at hand or hypothesis that φ, (2) the reliability of the source, and (3) their giving testimony as to that hypothesis. These variables and their relationships can be helpfully graphically represented in so-called Bayesian networks, directed acyclical graphs (DAGs) in which the nodes represent variables and the edges conditional dependencies:[9]

> "In [a] model with explicit source reliability, the quantity $P(E|H)$, the so-called likelihood, is replaced by the quantity $P(E - REP|H, REL)$, that is, the probability of an evidence report given that the hypothesis is true and the source is reliable. In other words, the likelihood of an evidence report is now a function of both the hypothesis and the reliability of the source. However, the reliability variable can be eliminated through marginalization in order to specify the probability of the evidence report conditional on H only[.] In other words, the explicit model can be reduced to the basic model [...] and, in this sense, the combination of both source and message content, explicitly considered, is in fact just another likelihood ratio." (Hahn, Harris, and Corner 2009, p. 346 f.)

9. For a brief introduction to Bayesian networks, see chapter 3.5 of Bovens and Hartmann 2004.

Trivially, providing just any DAG, dubbing one of its nodes 'source-reliability', initializing it with a probability distribution which obeys the probability axioms, and updating it using Bayesian conditionalization does not conclude the philosophical work. Instead, any source-reliability model must at minimum be partly reflective of the underlying 'causal structure' of the testimonial situation. Only if it is, can the offered systematic way of calculating the likelihood ratio based on the credences the Bayesian reasoner places in the hypothesis and their source's degree of interest alignment, competency or reliability simpliciter, be regarded as useful. While all of the source-reliability models considered in this thesis do pass this bar, they all differ in how much of the testimonial situation they are able to accurately reflect, as will become clear upon their introduction. Additionally, simple as they are containing only three to four nodes, they all come with further conceptual or 'cognitive' limitations that I will address throughout the thesis.

Applying these Bayesian source-reliability models has been adequately dubbed *expectation-based updating*, as opposed to *outcome-based updating*, which instead "requires definite knowledge on the actual outcomes" (Merdes, Sydow, and Hahn 2021, p. S5782, see also Hahn, Merdes, and Sydow 2018): based on an agent's priors about the reliability of their advisor and the truth of a proposition under consideration, they will form an expectation about the type of testimony they might receive. If you are entirely uncertain which door to pass through, you might have some trouble forming specific expectations about what any type of gargoyle would advise. However, if you already entertain a credence of 0.9 that it is the door to the left and think it is highly likely that the gargoyle is perfectly reliable, then you should expect them to point you to that very door with high probability. Next, upon actually receiving testimony from your advisor, expectation-based updating means comparing the expected to the actual report, and from that, drawing conclusions about source-reliability and the hypothesis. Assuming the gargoyle indeed tells you to enter the left door, then the fact that they confirmed your expectation might lead you to increase both your credence in their reliability and in the truth of the hypothesis. If, on the other hand, they unexpectedly tell you to enter the right door instead, this mismatch might weaken your belief in their reliability. Owning to your initial trust, however, you should still become less certain in your conviction to enter the left door.

Reading through this informal example, you might immediately suspect this form of belief revision to be crucially flawed. After all, an agent performing expectation-based updating changes their beliefs not based on observed and confirmed outcomes, but on whether incoming reports confirm their preexisting beliefs. They do not, for instance, open up both doors and re-evaluate the gargoyle's reliability based on the thus

confirmed truth-value of their advice. Given that the Bayesian machinery is largely silent on the question of which prior beliefs a reasoner should entertain, as long as they obey the constraints posed by the probability axioms, it is clearly compatible with beliefs that would be disconfirmed upon receiving correct advice. Subsequently, expectation-based updating may lead Bayesian reasoners starting with incorrect assumptions astray. This is another example of what Hájek and Hartmann (2010, p. 100, see also Assaad 2022, p. 105) call the 'garbage in, garbage out' problem. While "Bayesianism [may] play a salutary role in keeping our degrees of belief, like our beliefs, in harmony, and in policing our elicit inferences[,] the Bayesian constraints on degrees of belief should not be regarded as complete[, and] some additional constraints may well find their inspiration in traditional epistemology". However, just like Goldman, Hájek and Hartmann remind us that 'garbage in, garbage out' also features in deductive logic, wherein valid arguments may lead to incorrect conclusions if built on incorrect premises. As such, it would be premature to discard Bayesian, expectation-based updating solely on these grounds.

3.2 Credence density functions

Recall that the initial formulation of the gargoyle example only featured the most extreme, binary cases on the advisor spectrum. Some of the Bayesian source-reliability models that I will introduce in this chapter mirror this simplification via the limited way in which they conceptualize source-reliability: because they represent source-reliability using Boolean nodes in their DAGs, agents reasoning with these models may entertain fine-grained credences about whether or not a source is reliable, but fundamentally, they can only conceptualize a source as perfectly reliable, perfectly anti-reliable, or as a perfect randomizer. For example, though the details might only become completely clear once these models are formally introduced, assume that an agent reasoning using a Boolean source-reliability model places a credence of 60% in the reliability of their advisor. This does not mean that they think them 60% reliable, but rather, that they assign a subjective probability of 60% to 'the advisor is perfectly reliable'. In Figure 1, they could represent the axes and (some of) the extreme points marked by dots, but nothing in between.

Boolean source nodes also give rise to a second, related issue: reasoning in this manner, agents are incapable of reflecting how much information has previously gone into their belief formation. An agent with a randomly initialized source-reliability prior is indistinguishable from an agent that arrived at that same credence after collecting and updating on vast amounts of evidence. This is intuitively problematic if such

agents are to be used to model repeated interactions and the continuous gathering of evidence in social simulations, as we would expect agents to become more steadfast in their beliefs as they receive additional reports to support them:

> "[I]f we have relied on someone for a longer period of time, finding the reports of that person regularly confirmed, [they have] in our eyes built up a considerable track record as an informant. Given the considerable weight of the evidence favoring trust, a few misfortunes may not significantly alter the trust we place in [them]." (Vallinder and Olsson 2014, p. 2003)

However, there is no way for a model with Boolean source-reliability nodes to represent such changes in steadfastness without also altering the expectation of source-reliability: agents using these models can only become steadfast by updating their credences to become more extreme (closer to 0 or 1). For example, an agent believing that their advisor is 99.99% reliable will indeed take more time to be convinced otherwise than if they believed them 55% reliable instead, but it is impossible for such an agent to keep their expectation of reliability fixed while varying their steadfastness.

The remedy for both problems is initializing source-reliability models with continuous sources nodes, over which Bayesian agents may entertain distributions of second-order credences: in a nutshell, this allows them to entertain fine-grained beliefs about the exact degree to which their advisor is (anti-)reliable (see also Section 5.2), and the shape of the distribution contains enough information to distinguish between different amounts of evidence, as their variances decrease with additional information (see also Section 5.4). The original Bayesian model of source-reliability introduced to social epistemology by Bovens and Hartmann in 2004 (see Section 3.3) features a Boolean reliability node and thus displays this limitation, whereas the Bayesian agent introduced by Olsson (see Section 3.4) avoids the problem by representing a source's reliability with a continuous node. I will follow these leads and propose both Boolean and continuous versions of every focal Bayesian model I present in this chapter. The remainder of this section introduces the technical details behind (updating) credence density distributions.[10]

Representing source-reliability with a Boolean node suggests that it is ultimately just a lack of information that stops agents using the model from ditching any credence $\notin \{0, 1\}$ for one of the 'exactly two underlying possibilities'. Uncertainty when

10. Technically, it might also be interesting to have agents entertain credence distributions about the hypothesis so that they can be more and less steadfast with respect to that. However, just like Goldman and Duijf, I am interested in hypotheses concerning Boolean propositions, which would be misrepresented by continuous source nodes.

modelling more than two possibilities is instead represented by a distribution over the whole spectrum of possible probability values. With perfect access to information, an agent reasoning in this way would sharpen their distribution as such:

> "Let us suppose that there is a true probability P assigned to some proposition. We might also imagine having a subjective second-order probability distribution Q(.) over the possible values of P. Here, we shall make the simplifying assumption that P can take on a finite number of different values, for example, 0.01, 0.02 . . . . . 1.00. We treat P as an unknown parameter, about which we have subjective beliefs just as if it were any other unknown parameter such as the population of New York City. If we are very uncertain about the true value of P, Q(.) will spread out. If, on the other hand, we are certain of P, Q(.) will be concentrated at a point: in the case of a fair coin, Q(P) might be zero for all values of P except 0.50, where Q(P) would be 1." (see Baron 1987, p. 26 f.)

An agent entertaining a simple prior credence in a Boolean hypothesis and a prior credence distribution over the many source-reliability possibilities ultimately seeks to adjust the former to be close to the actual Boolean value of the hypothesis and to concentrate the latter as narrowly as possible around the actual reliability value of the source. Of course, in 'continuous models' there are infinitely many values these source nodes could take on, but still only a limited amount of 'probability mud' to distribute over them. Hence, second-order credences are represented not by assigning credence values to exact probabilities, but instead by so-called credence density distributions.

A common and natural way of initializing their priors is via the 'beta function', that is, using the following formula:

$$\frac{x^{\alpha-1}\overline{x}^{\beta-1}}{\frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}}$$

The shape of the distribution is determined by the numerator $x^{\alpha-1}\overline{x}^{\beta-1}$, whereas the denominator $\frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$ serves to normalize the function so that the total area under the curve (in the unit interval) sums up to one so that it can serve to represent a rational credence distribution. The resulting so-called 'beta distributions' have two features which make them particularly useful to the Bayesian enterprise. Firstly, they are robust to standard Bayesian updating on the outcomes of Bernoulli trials, random experiments with binary outcome possibilities: if one uses Bayes' formula to update a beta distribution about, for example, the chance embedded in a coin upon observing it being flipped repeatedly, one always receives another beta distribution back. Secondly, it is extremely simple to calculate the expected value of a beta distribution, which is given by $\frac{\alpha}{\alpha+\beta}$. Figure 2 shows two examples of beta-distributions being updated and

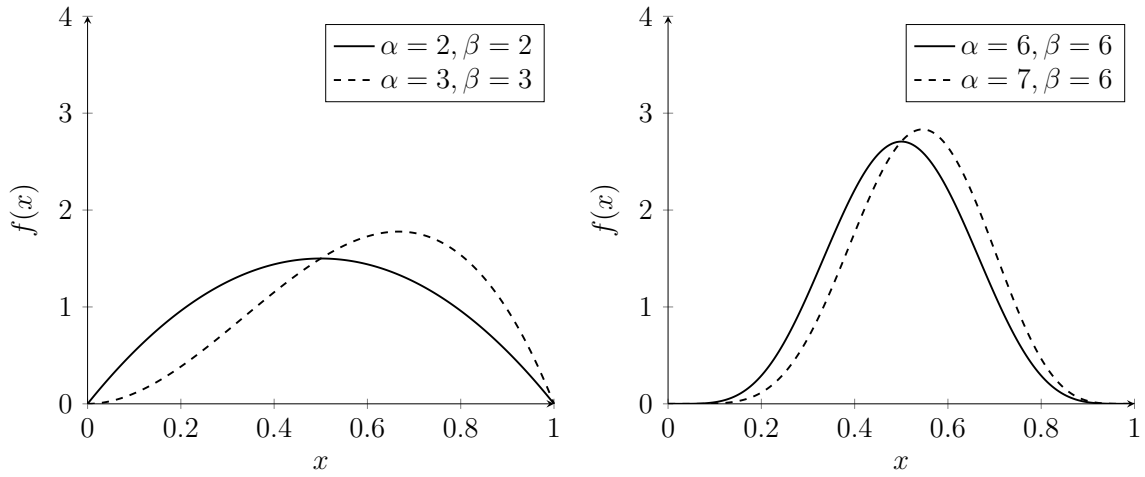

Figure 2: Example of how steadfastness can differ between two credence density functions with the same expected value. While both functions start out (continuous lines) with an expected value of 0.5, incrementing α by one (dashed lines) leads to a new expected value of 0.6 on the left, and of only roughly 54% on the right.

how they differ in their steadfastness depending on the α and β values involved.

To update second-order credences about a hypothesis H upon receiving a piece of evidence E, we can use the following formula:

$$h(x|E) = \frac{p(E|H = x)h(x)}{p(E)} = \frac{p(E|H)x + p(E|\neg H)\overline{x}}{p(E)}h(x)$$

Here x is a probability value in the unit interval, $h(x)$ the credence density assigned by the function h, and to calculate the unconditional $p(E)$ in this formula we may use the mean or expected value of h.

As an example, imagine a Bayesian reasoner entertaining second-order credences in a hypothesis H. The posterior $h(x)$ of their belief function is then given by:

$$\frac{p(E|H)x + p(E|\neg H)\overline{x}}{p(E|H)M(h) + p(E|\neg H)\overline{M(h)}}h(x)$$

Assuming the simple case of $p(E|H) = 1, p(E|\neg H) = 0$, this becomes $\frac{x}{M(h)}h(x)$ instead. If in this example, their priors are described by the neutral beta distribution obtained by setting $\alpha = \beta = 2$. Then the expected credence in the hypothesis ($M(h)$) is simply given by the formula $\frac{\alpha}{\alpha + \beta} = 0.5$. Let us calculate these posteriors for just two values of x: for $x = 0.1$, and this choice of beta-distribution, $h(x)$ decreases from 0.54 to 0.108, whereas for $x = 0.9$ the posterior rises instead, from 0.54 to 0.972. As

expected, observing evidence in favour of the hypothesis decreased our second-order credence in the low hypothesis probability 0.1, and increased our second-order credence in the higher H-probability 0.9. Helpfully, these posteriors perfectly match the values of $h(x)$ for the beta distribution with $\alpha = 3, \beta = 2$. For standard Bayesian updating, we can directly encode the number of hypothesis-confirming E in α (and of $\neg E$ in β), significantly simplifying the updating procedure for the distribution.

This is particularly useful when modelling large groups of agents repeatedly performing Bayesian inquiry, as observed when comparing two famous models by Zollman (2007, 2010). In the earlier version, the inquiry of individual agents is merely concerned with distinguishing between two states of a Boolean hypothesis and they entertain only simple credences. In the latter, they entertain second-order credences which get initialized with priors described by beta-distributions. However, Zollman is able to keep the overall computational complexity in check despite this, as updating these distributions and calculation of their expected values can simply be handled by incrementing α, β as needed and applying the formula $\frac{\alpha}{\alpha + \beta}$.

However, for the updating procedures required by the various Bayesian source-reliability models about to be introduced, this helpful effect does not persist. While one may still initialize such agents with priors that perfectly match beta distributions, the distribution may have changed considerably even after just a single round. In computationally implementing these models, compromises must thus be made: the exact distribution over the infinitely many probability values in the unit interval can be approximated by choosing a discrete resolution for the unit interval and storing the finite number of values required to represent that. The expected values of probability density distributions may then be calculated via the probabilistically weighted mean $\sum p(x)x$, or via integrating the probabilistically weighted distribution $\int_0^1 p(x)xdx$.

3.3 Bovens & Hartmann's model

As the first Bayesian source-reliability model, the BH-model was introduced to social epistemology by Bovens and Hartmann (2004, chapter 4). It features a single, Boolean source-reliability node and effectively captures the perspective of a reasoner trying to distinguish between an advisor being perfectly reliable, or simply determining their testimony by flipping a (weighted) coin. See Figure 3 for its definition, Figures 4 and 5 for an overview of how much of the possibility space for the testimonial situation BH-agents can conceptualize, and Appendix-section 7.3 for the technical details of belief revision with the BH-model. The behaviour of the Boolean BH-agent has been

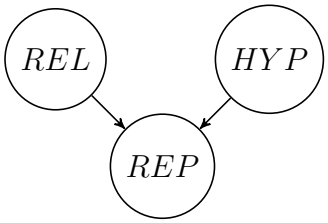

Figure 3: BH-model: the nodes are defined as follows: $p(HYP) = h, p(REL) = r$, and given the source is reliable, a report will occur iff the hypothesis is true, given the source is unreliable a report will occur with a frequency equal to the randomization parameter β:
$p(REP|HYP, REL) = 1$
$p(REP|\neg HYP, REL) = 0$
$p(REP|HYP, \neg REL) = \beta$
$p(REP|\neg HYP, \neg REL) = \beta$

much discussed in the literature already, so I will point the reader to Merdes, Sydow, and Hahn 2021 and Assaad 2022 for more in-depth discussions. However, given that continuously defining the reliability node of BH-agents is a new addition, I will detail some of the resulting behaviour in Sections 5.3 and 5.4.

One thing to keep in mind with the BH-model, as well as with the other focal models I will introduce in the next three sections, is that they model only situations in which an advisor does actually give a report. This means that they do not incorporate the possibility of an advisor being unable or unwilling to speak. Sometimes this will already accurately reflect the target scenario, like when a witness that has observed a crime happening is summoned in front of the court to testify whether they recognize a specific suspect, or when remaining silent can be clearly counted towards either of the two accepted answers. Other times, this is an idealization of the model that can be ameliorated by embedding it in larger models or descriptions, reducing the scope of the source-reliability model specifically to those cases in which an advisor does end up speaking, and having all other cases accounted for externally.

Before heading on to the next source-reliability model, let me provide an example of how the BH-model has been applied empirically: Bovens and Hartmann use it to attempt a rational reconstruction of Tversky and Kahneman's 'Linda-case experiments', to give an explanation of why participants seemed to be committing the so-called 'conjunction fallacy'. In a nutshell, the experiment asked participants what they judge to be more likely for a woman named Linda, who is 31 years old, single, outspoken, very bright, majored in philosophy, cared about discrimination and social justice, and demonstrated against nuclear-power plants: (A) Linda is a bank teller, or (B) she is

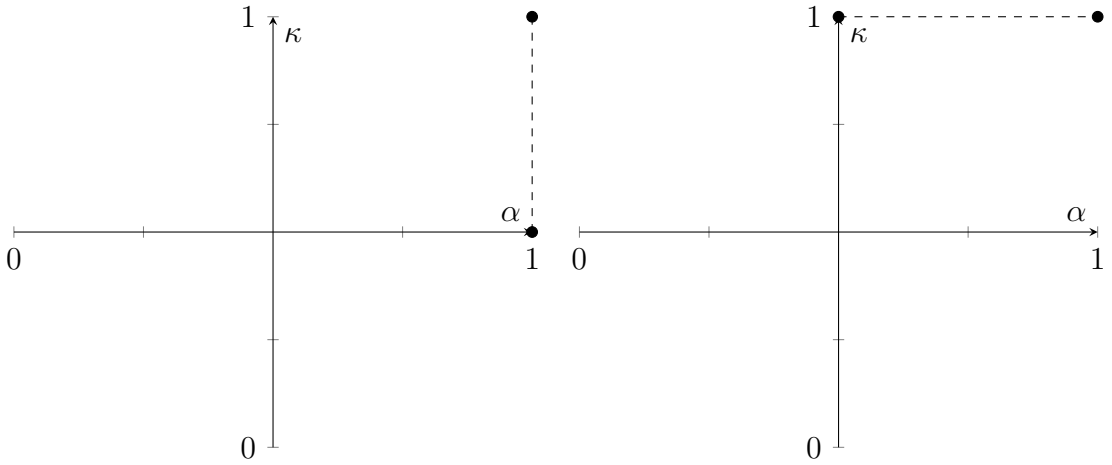

Figure 4: BH-agent with $p(REL)$ mapped to κ (left) and α (right). The Boolean BH-agent on the left can only fundamentally conceptualize an advisor as perfectly aligned but may either think them perfectly competent or a competency-based randomizer. The agent on the right can only conceptualize an advisor as perfectly competent but considers both the perfectly aligned variety, as well as alignment-based randomizers. The respective, continuous BH-agents can conceptualize what their Boolean counterparts do, plus everything in-between.

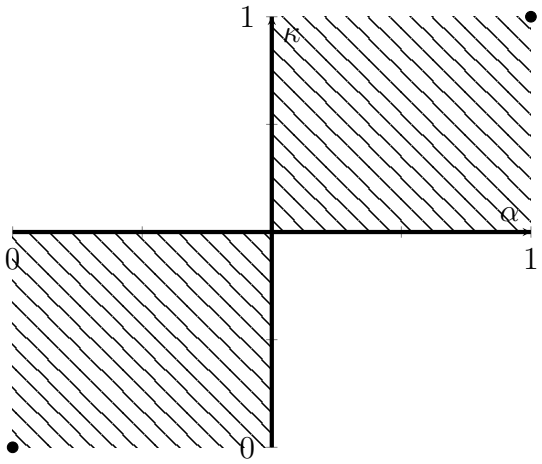

Figure 5: BH-agent with $p(REL)$ encoding reliability directly ($\kappa\alpha+\overline{\kappa\alpha}$). The Boolean BH-agent now captures both axes completely, as well as the extreme points of perfect reliability, and perfect double-anti-reliability. The continuous BH-agent can now fully conceptualize the first and third quadrants.

a bank-teller who is active in the feminist movement. The majority of respondents selected (B) as the more likely option, seemingly violating the above-mentioned probability axioms because every instance of (B) is necessarily also an instance of (A), and thus $p(A) \geq p(B)$.

Bovens and Hartmann use an extended version of the simple Boolean BH-model to show that a rational Bayesian reasoner entertaining the possibility that (A) or (B) is uttered by an unreliable source, may end up with a higher posterior belief in (B) upon receiving testimony that (B), than in (A) upon receiving testimony that (A). If one takes the background information about Linda provided by the experimenter as a given, then her being a feminist activist appears substantially more plausible than her being a bank teller. Now Bovens and Hartmann contrast (A) and (B) as two cases in which a potentially unreliable source[11] informs the recipient about Linda. Given their background knowledge, a participant receiving the unexpected testimony of just (A) may ascribe it to the possibility that their source is unreliable and simply happened to randomly affirm that Linda is a bank teller. Upon instead receiving testimony that (B), however, the expected assertion that Linda is a feminist may warrant the participant increasing their source-reliability credence, which will subsequently increase the impacts of the testimony that Linda is a bank teller. As a result, this recipient would barely increase their credence in case one compared to case two, allowing a slight reframing of the conjunction fallacy: instead of answering which statement (A) or (B) they think more likely, one might speculate that participants were actually answering upon hearing which of these statements their respective posterior degrees of belief in it would be higher.

However, while Bovens and Hartmann analysis matches the results of standard conjunction fallacy tasks, the picture is less clear as we look at its further predictions: we should expect adding another piece of information to (B), e.g., that Linda owns a copy of the Communist Manifesto, to increase the effects observed by Tversky and Kahneman 1983. Yet, these additional predictions have since been experimentally tested by Jarvstad and Hahn (2011), concluding that while the BH-model "fit the data reasonably well [...], other arguably more parsimonious models fit the data better" (p. 3039), and that specifically its "prediction that the addition of an extra component in classical conjunction problems would affect the incidence of the conjunction fallacy was not confirmed" (p. 3038).

11. The priors they chose specifically reflect the expectation of a 0.75 probability of correct advice, see also Section 4.1.

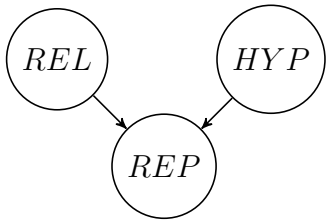

Figure 6: *OL-Model:* The nodes for the Boolean version of this model—which features a single form of anti-reliability—are defined as follows: $p(HYP) = h, p(REL) = r$, and given the source is reliable, a report will occur iff the hypothesis is true, given the source is unreliable a report will occur iff the hypothesis is false. The original version of the OL-model features a continuously defined reliability node instead.
$p(REP|HYP, REL) = 1$
$p(REP|\neg HYP, REL) = 0$
$p(REP|HYP, \neg REL) = 0$
$p(REP|\neg HYP, \neg REL) = 1$

3.4 Erik J. Olsson's model

This model comes originally embedded in *Laputa* (see Olsson 2011, 2013, Angere 2010), a modelling software in which every agent is already part of a social network. Various additional variables influence agents' communication in Laputa: agents perform their own inquiry governed by their 'aptitudes' and update on the results depending on their 'self-trust'; whether reports are even received by other agents depends on the 'listen chances' of an advisor's network-neighbours; and whether advisors testify that the hypothesis holds depends, apart from their $p(HYP)$, on a so-called 'threshold of assertion'.[12]

To get at the core of Olsson source-reliability model, however, I followed Merdes, Sydow, and Hahn (2021) in extracting it from its natural habitat and giving it a description in the same format as the other focal models in this section (see Figure 6). Appendix-section 7.6 describes the updating procedures for both versions of the OL-model, and Figure 7 describes which parts of the testimonial situations they are able to conceptualize.

Similarly to BH-agents, OL-agents possess only a single reliability node to store information about an advisor, but unlike them, they are much less optimistic: for OL-agents, an advisor may be anti-reliable, worse than chance.

Due to its symmetric nature, the Boolean OL-model specifically is deeply flawed:

12. This assertion threshold t distinguish systematically honest ($t > 0.5$), systematically lying ($t < 0.5$) and perfectly randomizing ($t = 0.5$) advisors: in the first case, they assert that the hypothesis is true only if $p(HYP) \geq t$, and that it is false only if $p(HYP) \leq \bar{t}$; in the second case they assert that the hypothesis is true only if $p(HYP) \leq t$ and that it is false only if $p(HYP) \geq \bar{t}$ (Angere 2010, p. 4 f.).

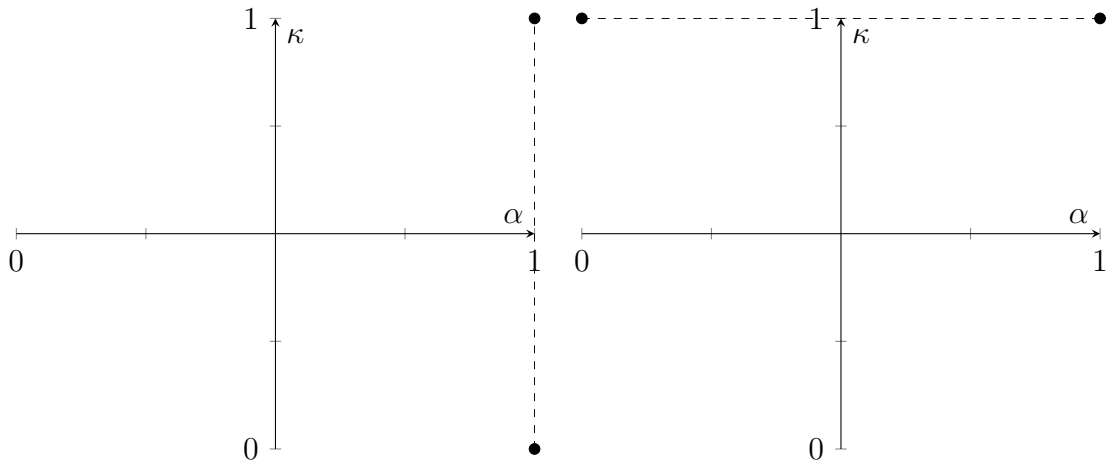

Figure 7: OL-agent with $p(REL)$ mapped to κ (left) or α (right). The agent on the left can only conceptualize an advisor as perfectly aligned but considers both the perfectly competent, and perfectly anti-competent extremes. The continuous OL-agent can conceptualize this plus everything in-between. The agent on the right can only conceptualize an advisor as perfectly competent but considers both the perfectly aligned, and perfectly anti-aligned extremes. The continuous OL-agent can conceptualize this plus everything in-between. If the reliability-node is taken to capture both competency and alignment, the OL-agent becomes capable of conceptualizing the complete range of possible advisors, akin to the SD-agent (see Figure 11).

for any combination of $p(HYP), p(REL)$ there is always a 'testimony path'—of at most length two—to $p(HYP) = p(REL) = 0.5$. Receiving a positive report first will set them to the same value, and then a single negative report sets them both to 0.5, where they subsequently get stuck. Alternatively, receiving a negative report first will result in posterior $p(REL), p(HYP)$ at an equal distance on either side of 0.5, which when followed by a positive report collapses into $p(REL) = p(HYP) = 0.5$.

3.5 Leon Assaad's 'Alignment model'

Introduced in Assaad 2022, and inspired by both Bovens and Hartmann 2004 and Duijf 2021, this Bayesian model of source-reliability disentangles the notions of source-alignment and source-competency into two separate Boolean[13] nodes (see Figure 8 for its definition, and Appendix-section 7.4 for its updating mechanisms). For an in-depth analysis of the behaviour of Boolean AL-agent, as well as for some results concerning the simulation of the interactions in communities thereof, see Assaad 2022.

The Boolean AL-model is almost identical to another Bayesian source-reliability

13. Given that the AL-agent uses two nodes to capture reliability, the option arises of defining only one of them continuously. I have detailed the conceptual space of such models in Appendix-section 7.2, but will otherwise not further consider such agents here.

model from the psychological science literature on argumentation: independently introduced by A. J. Harris et al. (2016, p. 1503f), it, in turn, appears to have been inspired by the BH-model and a model of epistemic trust introduced in Shafto et al. 2012. Additionally, it is a simplification of a more complex Bayesian source-reliability model introduced in Hahn, Oaksford, and Harris 2013, which itself is based on six questions about source-reliability taken from Walton 1997. A. J. Harris et al. label their competency-node 'expertise' and their alignment node 'trustworthiness', and for reasons of model simplicity they do not use a randomization parameter β, instead fixing the probability of both positive and negative reports from completely unreliable sources at 0.5. Their paper also features the results of two experiments which they take to empirically confirm the predictions of this model.

In accordance with the restrictions placed on the OD-model in Duijf 2021, AL-agents do not feature competency-based anti-reliability, as advisors may at worst be randomizers in that regard, see Figure 9 for an overview diagram. A side-effect of this restricted range of the COM-node is that AL-agents tend to be much more willing to adjust their credence about an advisor's alignment, than about their competency. Another way to express this is that because the AL-agent only considers alignment-based anti-reliability, they will more easily accredit unexpected reports to their advisor being anti-aligned, than to them merely being a competency-based randomizer, which has them quickly suspect that imperfect advisors may be lying to them.

As Assaad points out, if $p(AL)$ is set to one, the AL-model becomes mathematically equivalent to the BH-model. However, to maintain an accurate interpretation of what is going on here, keep in mind that by default, we should assume BH-agents to conceptualize the possibility space for advisors as depicted in Figure 5: with their reliability node capturing both competency and alignment. AL-agents with their alignment nodes set to one are better identified with BH-agents whose reliability node is mapped to κ (see Figure 4). While for $\alpha = 1, \alpha\kappa + \overline{\alpha\kappa}$ does equal κ, the relevant difference lies with the possibility of reliability due to simultaneous anti-alignment and anti-competency: by default, BH-agents might conceive of advisors as located anywhere in the first or third quadrant, whereas AL-agents with $p(AL) = 1$ are restricted to just a sliver of the first quadrant.

Lastly, read in the way I interpret competency and interest alignment in Section 2.2, the baseline AL-agent suffers from a form of internal inconsistency: as they assume the possibility of alignment-based anti-reliability, then qua the existence of testimonial chains like described in Section 2.1, they should also consider the possibility of competency-based anti-reliability. The fact, that they do not, may become problematic when AL-agents are placed in network simulations as Assaad himself does.

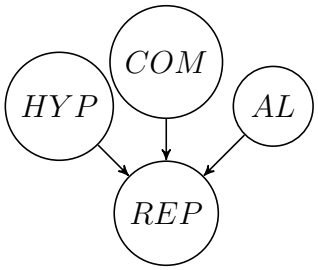

Figure 8: AL-model: the nodes are defined as follows: $p(HYP) = h, p(COM) = c, p(AL) = a$, and given the source is competent, a report will occur iff both or neither of the other two root nodes hold, given the source is incompetent, a report will occur with the frequency given by the randomization parameter β. To obtain the definition of $AL\star$ instead, simply switch out COM and AL.
$p(REP|HYP, COM, AL) = 1$
$p(REP|\neg HYP, COM, AL) = 0$
$p(REP|HYP, COM, \neg AL) = 0$
$p(REP|\neg HYP, COM, \neg AL) = 1$
$p(REP|HYP, \neg COM, AL) = \beta$
$p(REP|\neg HYP, \neg COM, AL) = \beta$
$p(REP|HYP, \neg COM, \neg AL) = \beta$
$p(REP|\neg HYP, \neg COM, \neg AL) = \beta$

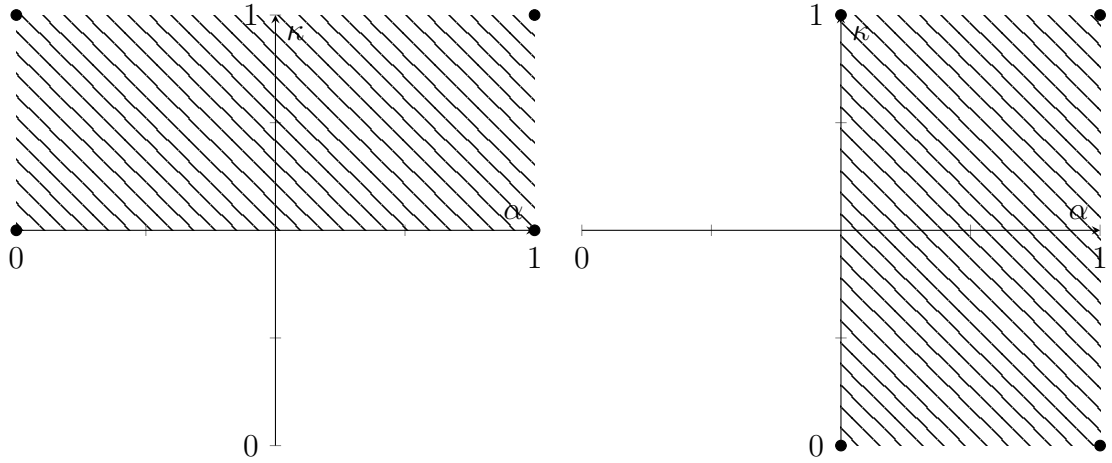

Figure 9: The Boolean AL-agent (left) can only conceptualize a source as the four combinations of fully aligned or fully misaligned, and of perfectly competent, or a competency-based randomizer. The continuous AL-agent (pattern) can conceptualize these four cases and every case in between them. The figure on the right shows what an alternative version $AL\star$ would capture: here, the agent can only conceptualize a source as the four combinations of fully competent or fully anti-competent, and of perfectly aligned, or as an alignment-based randomizer.

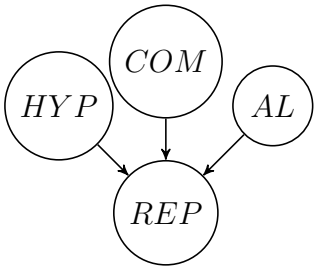

Figure 10: SD-model: the nodes are defined as follows: $p(HYP) = h, p(COM) = c, p(AL) = a$, and $p(REP) = 1$ iff an unequal number among HYP, COM, AL are equal to 1 (or, Given that one of the root-nodes holds, then a report will occur iff either both other nodes hold, or neither hold.):
$p(REP|HYP, COM, AL) = 1$
$p(REP|\neg HYP, COM, AL) = 0$
$p(REP|HYP, \neg COM, AL) = 0$
$p(REP|HYP, COM, \neg AL) = 0$
$p(REP|\neg HYP, \neg COM, AL) = 1$
$p(REP|\neg HYP, COM, \neg AL) = 1$
$p(REP|HYP, \neg COM, \neg AL) = 1$
$p(REP|\neg HYP, \neg COM, \neg AL) = 0$

Upon identifying a source as anti-reliable, AL-agents conclude anti-alignment, even in cases where they are being advised by a gullible, that is incompetent advisor who is being lied to by third parties. A different version of this focal model, $AL\star$ (see Figure 9) avoids this contradictory nature: $AL\star$-agents recognize the possibility of the full spectrum of competence-values, including competency-based anti-reliability. However, somewhat optimistically, $AL\star$-agents do not conceptualize the possibility of alignment-based anti-reliability: at worst they may suspect an advisor to randomize their testimony, but never invert it. While this optimistic assumption might be justified only rarely in practice, it is at least possible in principle that a combination of a domain of inquiry and a homogeneous population of agents conform to it.

3.6 The 'subjective Duijf model'

I created this specifically to mirror the OD-model from the viewpoint of the subjective perspective. Thus, as you can also gather from Figures 10 and 11, it can fully conceptualize the advisor spectrum based on using separate source nodes for alignment and competency. See Appendix-section 7.5 for the details about its belief revision process.

Here are four quick notes about this model: first, analogously to the relationship between the AL-model and the BH-model, setting $p(AL) = 1$ results in the SD-

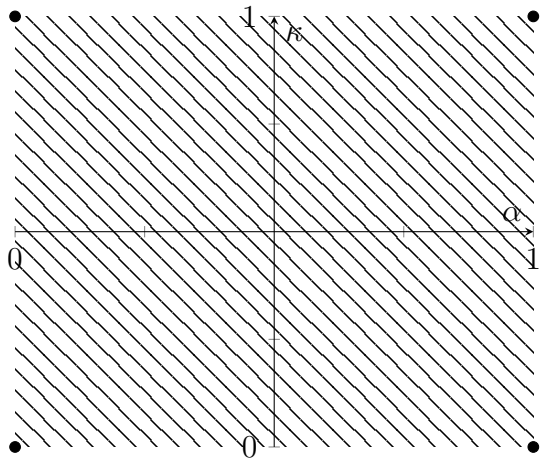

Figure 11: The Boolean SD-agent can only conceptualize an advisor as one of the four extreme combinations (black dots) of fully aligned or fully misaligned, and of perfectly competent, or perfectly anti-competent. The continuous SD-agent can conceptualize the entire possibility space (hatched pattern) for advisors as given by the objective model.

model becoming mathematically equivalent to the OL-model[14]. However, while the default OL-agent can fully conceptualize the entire advisor spectrum (see Figure 11, SD-agents with their alignment node set to 1 are better modelled as OL-agents with their reliability node mapped to κ (see Figure 7). Secondly, as soon as any source node takes on value 0.5, the model expects a report confirming the hypothesis with probability 0.5. This, in turn, is to be expected: if an SD-agent has no clue whether the hypothesis holds, or assumes that a source acts as a perfect randomizer about whether to give a positive report due to being neutral in terms of either competency or alignment, then they will stay neutral with respect to the kind of report they will receive. Thirdly, unlike the BH-model and AL-model, the SD-model does not feature a randomization parameter β. Instead of explicitly encoding the probability of a report given a randomizing source, it implicitly reasons as if that value was 0.5. Lastly, unlike the AL-model, SD agents are not more steadfast with respect to their competency credence than to their alignment credence: given that both nodes span equal distances of the advisor spectrum, unexpected reports will, ceteris paribus, be equally attributed to both reliability related source nodes.

14. To understand why, take a look at the updating procedures for both agents in Appendix-sections 7.5 and 7.6: setting $a = 1$ in any formula for the SD-model will render the respective formula of the OL-model.

3.7 Other (approaches to) focal models

Clearly, the source-reliability models just introduced do not even get close to exhausting all possibilities. In this section, I want to briefly address three more potential ways of tackling source-reliability estimation from the subjective perspective.

First, it is technically possible to directly apply the normative elements of Duijf's objective model from the subjective perspective, if one already holds full beliefs in all three relevant values, or is explicitly aware of the credences they place in them. Do you believe your advisor to be more reliable than you do yourself? In that case, you should defer to their advice or testimony. Do you at least believe them more reliable than chance? Then you might still want to listen to and update on their testimony. If a recipient holds a full belief that their advisor is competent to 80% with respect to an issue at hand, that they themselves are competent to only 60%, and that their interests align to 90%, then focally applying Duijf's model prescribes they defer to any advice received. The source-reliability models discussed in this thesis do not entertain beliefs or credences in their own reliability, but this issue can easily be circumvented. For example, agents in Laputa apply expectation based updating to both the reliability of their advisors, and to their own reliability as inquirers.[15] Of course, to apply the OD-model $1:1$ such agents would first have to ditch the subjective cntinuous OL-model for the more fine-grained continuous SD-model, which differentiate between aligment and competency. Alternatively, see Heinzelmann and Hartmann 2022 for an example of a Bayesian source-reliability model that encodes a node for self-trust directly in the DAG: the authors use it to present an account of how Bayesian reasoners should update their credences after asserting a hypothesis to others and not receiving any push-back.

Secondly, I want to quickly mention another focal source-reliability model that I will not be discussing further for the remainder of the thesis. Merdes, Sydow, and Hahn (2021, p. S5794) suggest extending the Boolean BH-model by adding a third possible value to its reliability node: the resulting model allows agents to conceptualize fully reliable sources, perfect randomizers, and fully anti-reliable sources (see Figure 12). Unlike the Boolean OL-model and SD-model, this trivalent BH-version is at least somewhat capable of handling a set of inconsistent reports, which it takes as definitive proof of a purely randomizing advisor (see Section 4.4). However, ultimately it still misconceptualizes the testimonial situation and will at the very least be misled about advisors that randomize to any other degree than, and can thus be safely neglected.

15. See Pallavicini 2018 and Olsson 2020 for a discussion of expectation-based updating of self-trust of agents as a separate source, that is, expectation-based updating of agents' trust in their individual inquiry, not own priors.

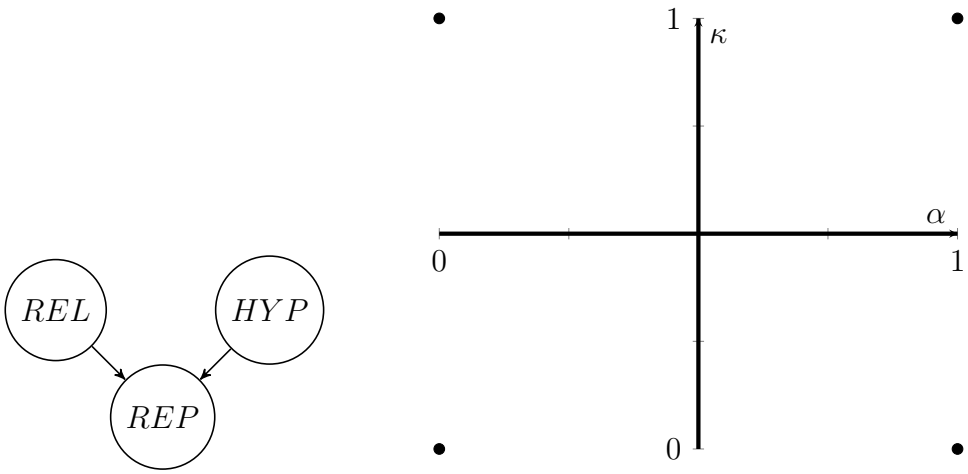

Figure 12: Trivalent BH-agents, as suggested by Merdes, Sydow, and Hahn 2021, could conceptualize an advisor as one of three special cases: perfectly reliable sources ($REL = 2$), perfectly anti-reliable sources ($REL = 0$), and randomizers ($REL = 1$). If their reliability node is taken to encode the probability of correct advice, this comes down to the four corner points and both axes of the advisor spectrum.

And thirdly, another avenue that might be fruitfully explored in future work: compared to the continuous versions of the BH-model and AL-model, the OL-model and SD-model each helpfully capture twice the amount of possibility space on the advisor spectrum. Sometimes, however, we might actively seek to restrict which possibilities a focal agent can identify an advisor as, like when we want to investigate the effects of very trusting listening practices, or when adapting the agent to a purely honest reporting context. Creating and implementing ever new models to exemplify exactly the currently desired criteria might not be the most cogent route. Instead, we might consider sticking to the same, unifying OL-model or SD-model, and outsource the fine-tuning to the initial credence distributions. For example, instead of switching from the OL-model to the BH-model, one could initialize the initial reliability credence distribution as equal to 0 within $[0, 0.5)$, and normalize it in the remaining half of the interval. We might interpret the use of such a model as equivalent to conceptually entertaining the possibility of the entire advisor spectrum while being simultaneously absolutely certain that some parts of it are not represented in the reporting context at hand. Alternatively, if expectation-based updating with the possibility of anti-alignment actually does cause polarization (see Olsson 2020b) or even deep disagreement (Assaad 2022), then if one suspects it to be sufficiently rare in the reporting contexts, one might want to restrict all credence density to the reliable quadrant(s) for pragmatic reasons.

4 Model evaluation: background

Before we are ready to computationally evaluate these focal Bayesian models, there remain a few preliminary issues to discuss. In the first section of this chapter, I detail how representing the same point on the advisor spectrum requires different value assignments for the different focal models, and how to translate between each of them and the OD-model. Subsequently, I argue for *trustworthiness accuracy* as a secondary measure for the adequacy of source-reliability agents, to go alongside the change in veritistic value caused by their application, using this opportunity to delve deeper into some conceptual details. Next, I address the issue of fine-tuning Bayesian source-reliability models—or their priors—to the reporting contexts in which they are applied, concluding that both should mostly be disregarded. In the fourth section, I explain the notions of global and local updating on sets of reports received from a single source. I argue that despite the in-applicability of global updating to (group) simulations, we can still use it as a benchmark for sorting out conceptually inadequate source-reliability models. Finally, Section 4.5 takes a closer look at group simulations like Laputa and identifies two crucial problems with their application of Bayesian source-reliability models.

4.1 Inter-model translation

David Lewis' 'Principal Principle'[16] (1980) tells us that rational agents if they have the required information, ought to conform their subjective probabilities to the underlying objective chances. This raises the question of how to translate from objective probability values encoded by the OD-model to the credences of our focal Bayesian agents—and back:

For some of the source nodes introduced in the last chapter, this process is rather straight-forward: the alignment nodes of the Boolean AL-model and Boolean SD-model, as well as the competency node of the SD-model, can simply be transferred $1:1$ as the respective values in the OD-model. The same is the case for the REL-node of the OL-model, iff it is mapped directly to competency or alignment. For example, to model an advisor with 60% competency and 70% interest alignment, a SD-agent wants their subjective $p(COM) = 0.6$ and $p(AL) = 0.7$. To initialize continuous versions of these models, one has to choose probability distributions with an equal expected value, choosing flatter or steeper distributions to represent less or more steadfast credences.

However, the reliability node of the BH-model, the competence node of the AL-

16. Also known as the law of direct probabilities.

focal agent	REL	COM	AL
Boolean BH-agent	$p = 0.36$	—	—
continuous BH-agent	$\alpha = 1, \beta = 1.\overline{7}$	—	—
Boolean OL-agent	$p = 0.68$	—	—
continuous OL-agent	$\alpha = 2.125, \beta = 1$	—	—
Boolean AL-agent	—	$p = 0.6$	$p = 0.8$
continuous AL-agent	—	$\alpha = 1.5, \beta = 1$	$\alpha = 4, \beta = 1$
Boolean SD-agent	—	$p = 0.8$	$p = 0.8$
continuous SD-agent	—	$\alpha = 4, \beta = 1$	$\alpha = 4, \beta = 1$

Table 1: Example of how the reliability-related priors of these focal models differ when representing the subjective expectation of $\kappa = \alpha = 0.8$. The α and β values chosen for the continuous agents are the smallest possible given the computational necessity that $\alpha \geq 1, \beta \geq 1$.

model and the alignment node of the $AL\star$-model each just range over half the possibilities compared to all the other source nodes: from the expectation of a perfectly accurate advisor to that of a purely randomizing one. As a result, not only are the credences associated with these nodes more steadfast than those associated with 'full' nodes, but they also need to be initialized differently to allow proper comparison: Where an SD-agent entertains a $p(COM) = 0.75$, an AL-agent instead believes that $p(COM) = 0.5$. For the objective probabilities ($0.5 \leq p \leq 1$) which these agents can conceptualize, we arrive at their respective credence c via $c = 2(p - 0.5)$.

When mapping the single reliability nodes of OL-agents and BH-agents to an advisor's alignment and competency simultaneously, we can use the same mechanisms but must replace κ, α with the probability of correct advice $\alpha\kappa + \overline{\alpha\kappa}$: for the OL-agent, we can adopt this probability $1 : 1$, for the BH-agent—iff it is between 0.5 and 1—we use the above formula. Note that when initialized in this way, both BH-agents and OL-agents will naturally—and correctly—treat doubly anti-reliable advisors as reliable sources.

Properly translating between these different focal models is often required to adequately compare them computationally: instead of comparing their behaviours when we initialize every node with the exact same priors, we should instead initialize them such that their priors each translate into the exact same objective chances in the OD-model. For instance, to compare how these focal agents update their credences when initialized in accordance with the expectation that $\kappa = \alpha = 0.8$, we ought to choose priors for each focal agent as listed in Table 1. For the second trust initialization setup used in Chapter 5, the barely trusting expectation of $\kappa = 0.6$ and $\alpha = 0.7$, see Table 2.

focal agent	*REL*	*COM*	*AL*
Boolean BH-agent	$p = 0.08$	—	—
continuous BH-agent	$\alpha = 1, \beta = 11.5$	—	—
Boolean OL-agent	$p = 0.54$	—	—
continuous OL-agent	$\alpha = 1.17391, \beta = 1$	—	—
Boolean AL-agent	—	$p = 0.2$	$p = 0.7$
continuous AL-agent	—	$\alpha = 1, \beta = 4$	$\alpha = 2.\overline{3}, \beta = 1$
Boolean SD-agent	—	$p = 0.6$	$p = 0.7$
continuous SD-agent	—	$\alpha = 1.5, \beta = 1$	$\alpha = 2.\overline{3}, \beta = 1$

Table 2: Initializing of reliability-related priors when representing the subjective expectation of $\kappa = 0.6, \alpha = 0.7$, using minimal α and β-values for the prior credence density distributions.

4.2 Trustworthiness accuracy

In Section 2.3, I introduced the possibility of evaluating focal source-reliability models as listening practices in terms of Goldman's veritism: how does the veritistic value of reasoners change over time when they use these models to update their beliefs? This evaluation is grounded in the fundamental value assigned to holding true beliefs about the hypothesis so that models that lead to increased veritistic values with respect to it earn epistemic praise, and those leading to decreasing veritistic value are judged problematic. Let us dub this form of evaluating Bayesian source-reliability models *veritistic accuracy*. In light of this thesis' main focus on source-reliability, however, another angle of evaluation opens up: how accurately do reasoners employing these focal source-reliability models assess the trustworthiness of their source, and how correct are these reliability estimates? Let us call this second form of evaluation *trustworthiness accuracy*.

Here are two reasons for considering trustworthiness accuracy a fruitful avenue: firstly, how reliable one's advisor is, poses an interesting question all on its own, one that a recipient can hold more or less accurate beliefs about. Secondly, it may let us zoom past confounding factors in the veritistic analysis and directly investigate source-reliability models at their core. The practice of trusting (or even deferring to) an advisor has the proclaimed aim to trust advisors iff they are trustworthy, i.e., sufficiently competent and aligned. Verisitstic accuracy, however, is directly dependent on the distribution of the reliability of the sources: no matter what, confronted by a purely randomizing advisor, a recipient will be unable to consistently improve their veritistic value about the hypothesis—at best they would not worsen it—but they may well correctly estimate source-reliability in the right circumstances. Additionally, the veritistic analysis depends on how often a recipient gets a chance to update, or the

extent to which they apply this model as an epistemic listening practice: updating only a single report, for everything but the most extreme reliability-priors, has a limited impact on veritistic value, whereas for updating on ten or more reports the potential is much greater. For veritistic analysis, it can be difficult to decide when to draw results.

There are two interesting angles from which we might approach the creation of a trustworthiness accuracy measure. Firstly, we may look to measure the verisimilitude of the reliability credences entertained by focal agents, essentially reapplying the idea of veritistic value to the non-Boolean issue of determining source-reliability. However, it would be quite difficult to produce a scoring rule in this spirit that does not outright discriminate against some of the models under consideration. It is unclear how to weigh issues (i) through (iv) in a fair verisimilitude measure without running the risk of tailoring it to specific models ad hoc:

(i) The Boolean versions among our models, as they fundamentally conceptualize any source as perfectly reliable, perfectly anti-reliable or full randomizers, could only receive perfect trustworthiness accuracy scores for these exact cases. For any intermittent reliability values, even accurate seeming credences would still express fundamentally incorrect beliefs and would have to be evaluated as such. For example, even if a Boolean OL-agent assigns a subjective $p(REL) = 0.7$ to a source that objectively gives correct advice with that probability, the agent conceptualizes them as either perfectly reliable or perfectly anti-reliable and is merely leaning towards the former.

(ii) Only agents reasoning with versions of the AL-model and SD-model are able to conceptually disentangle alignment and competency and assign subjective degrees of belief to both. This ought to count towards a verisimilitude score to at least some extent: correctly assessing an advisor's reliability qua correctly identifying their exact alignment and competency values and possessing the knowledge of how to combine them captures the testimonial situation to a much larger degree.

(iii) Even if agents accurately assess the probability of correct advice, problems do remain: in the vacuum of pure, expectation-based updating on Boolean testimony from a single source, it is indistinguishable whether a reliable source belongs to the first or third quadrant, i.e., whether they are both competent and well-aligned, or neither. Similarly, it is impossible to distinguish between anti-reliable sources from the second or fourth quadrant, i.e., between competency-based and alignment-based anti-reliability. For OL-agents and BH-agent, this distinction

is not represented in their coarse-grained reliability credences in the first place, whereas SD-agents and AL-agents will capture the distinction, but simply lack the information to decide for one over the other. Among the combination of competency and alignment that are compatible with the observed reliability, or frequency of agreeable reports, they will simply settle on those that are closest to their own priors.

(iv) I previously discussed as an upshot of using continuous source nodes, that the resulting agents can distinguish different amounts of certainty or information backing up one and the same mean degree of belief in competency, alignment or reliability simpliciter. To what extent should this capability count towards our verisimilitude measure? Take, for example, two continuous OL agents who correctly identify a source as slightly anti-reliable, each holding a mean reliability credence of 0.4. However, where one of these agents is extremely certain in their stance——their credence density distribution is given by the beta function with $\alpha = 200, \beta = 300$—the other is way less steadfast with $\alpha = 2, \beta = 3$.

As a second avenue, we may instead seek to make use of the normative component of the OD-model, and evaluate an agent's trustworthiness accuracy depending on how often they correctly follow its suggestions. That is, deferring to the testimony of an advisor if and only if their probability of giving correct advice $\alpha\kappa + \overline{\alpha}\overline{\kappa}$ is greater than one's own competency ρ. However, there are two immediate problems with applying this measure.

- Firstly, it is unclear how we should understand ρ in a single run of the computational model: where an advisor's alignment and competency values can be identified based on the frequency with which they generate (in)correct advice, there is only a single value that we can use to approximate the recipient competency. Recall the definition of a recipient's competency ρ as the probability that, presented with an arbitrary Boolean question φ from the relevant domain of inquiry, the recipient would correctly assess the truth value of φ. If we instead tasked the recipient with assigning a degree of belief in φ, then absent reasons to the contrary, and over multiple runs of the model, we should expect their initial veritistic values to average to their competency score. Still, identifying veritistic value with ρ is everything but a smooth fit, and choosing a measure of trustworthiness accuracy which cannot be properly applied to individual model-runs seems unreasonable.[17]

17. Alternatively, one might understand ρ dynamically as the current veritistic value at each round,

focal agent	**trusting iff**	**anti-updating iff**
Boolean BH-agent	$p(REL) > 0$	—
continuous BH-agent	$M(r) > 0$	—
Boolean OL-agent	$p(REL) > 0.5$	$p(REL) < 0.5$
continuous OL-agent	$M(r) > 0.5$	$M(r) < 0.5$
Boolean AL-agent	$p(AL) > 0.5, p(COM) > 0$	$p(AL) < 0.5, p(COM) > 0$
continuous AL-agent	$M(a) > 0.5, M(c) > 0$	$M(a) < 0.5, M(c) > 0$
Boolean SD-agent	$\frac{(p(AL)\ p(COM)+}{\overline{p(AL)}\ \overline{p(COM)})} > 0.5$	$\frac{(p(AL)\ p(COM)+}{\overline{p(AL)}\ \overline{p(COM)})} < 0.5$
continuous SD-agent	$\frac{(M(a)\ M(c)+}{\overline{M(a)}\ \overline{M(c)})} > 0.5$	$\frac{(M(a)\ M(c)+}{\overline{M(a)}\ \overline{M(c)})} < 0.5$

Table 3: For which credence values in the reliability, competency and alignment does each Bayesian source-reliability model update their hypothesis credence in line with incoming testimony? For continuously defined agents, the means $M(r), M(c), M(a)$ of the respective distributions are decisive instead.

- Secondly, agents reasoning with Bayesian source-reliability models do not necessarily ever properly 'defer': they never perform any actions that go beyond belief revision, and they never simply adopt their advisor's testimony $1:1$ either. While the model does contain a few features that could be used to approximate the meaning of deferral, this too appears ad hoc: For example, to count a recipient's belief revision in any individual round as deferral, we might require them to (i) update their hypothesis credence in the direction of the received testimony, (ii) increase their (mean) reliability credence for that advisor, and (iii) end the round with a $p(HYP)$ with a distance of < 0.5 from the advice received.

Having unsuccessfully explored these two avenues, of evaluating the verisimilitude of the recipients' source-reliability credences, and of deferring iff the OD-model prescribes it, I suggest an altogether simpler compromise between them: we should require recipients to update positively on advice given by advisors that are more reliable than chance, anti-update on testimony from anti-reliable advisors, and ignore all advice from randomizing sources. Table 3 shows for which (mean) credence values these focal agents put advisors in which of these three boxes.

Even without running any simulations, it should immediately become clear that we ought to expect substantial differences in trustworthiness-accuracy scores between agents utilising different Bayesian source-reliability models: for instance, BH-agents, incapable as they are to conceptualize a source as anti-reliable, will necessarily have

but see Section 4.5 for more details about how defining reliability for simulations can be exceptionally tricky.

0% trustworthiness accuracy whenever they interact with advisors belonging to the second or fourth quadrant. Conversely, when advised by those within the first or third quadrant, we should expect their trustworthiness-accuracy to be infallible (given their reliability credence isn't exactly equal to 0). Similar—if less clear-cut—reasoning applies to the remaining models.

4.3 (Against) fine-tuning

Rini (2021) provides a case study of how Russian state-sponsored social-media-based misinformation interfered with the functioning of democracy in the US between 2014 and 2018. Not only did this interference create false beliefs and helped them spread, but it also undermined the trust of citizens in each other, as the notion of advisors being paid liars, or even bots was present in their minds. When modelling cases like this one, it may be wise to make use of Bayesian source-reliability models which are quick to assume (specifically alignment-based) anti-reliability. Similarly, upon identifying a community of inquirers as predominantly honest and competent, it might be fruitful to ensure high expectations of reliability, for example by choosing a source-reliability model that cannot conceptualize advisors as anti-reliable. More generally, this kind of thinking suggests that to tackle the problem of rational source-reliablity estimation, we ought to collect empirical data on the distribution of competencies across the relevant potential advisors, as well as on how widespread alignment-based anti-reliability is among them. Subsequently, one might then wish to fine-tune or pick and choose only those Bayesian source-reliability models whose awareness of the advisor spectrum most closely resembles this data.

In a way, Goldman (1999, p. 109) anticipates such ideas, when he describes how message-acceptance strategies, like e.g., the blind trust a OL-agent with $p(REL) = 1$ would exhibit, depend in their veritistic effectiveness on the reporting environment in which they are applied. 'Blind trust' is an excellent strategy when used in a community of competent truth-tellers (or anti-competent liars, though this might be a rather rare find), but a bad strategy almost everywhere else. Similarly 'blind contratrust', quickly assuming anti-alignment and steadfastly trusting in advisors' competencies might be an excellent strategy in a community made up exclusively of competent liars.

> "Alternatively, social epistemology might take the *de facto* distribution of reporting practices as a given, for good or ill. An acceptance practice should be crafted that would complement the existing distribution of reporting practices. Since different reporting practices might be used in different contexts—one practice when reporting the weather and a different practice when declaring one's

> annual income to the revenue service—perhaps epistemology should advise hearers to vary their acceptance practices as a function of those different contexts. But should epistemology take on the assignment of trying to describe *de facto* reporting practice in every social niche and sector of discourse? From a practical point of view, this would be excessive. Even I, whose aspirations for epistemology might strike some as grandiose, am not prepared to ask *that* much of it." (Ibid., p. 110)

He goes on to suggest a 'satisficing approach': instead of fine-tuning acceptance practices to specific distributions of reporting strategies, to aim at finding a single acceptance practice that is good enough for all (relevant) reporting practices. In fact, this is the point at which he champions (restrained) Bayesian inference—introduced in Section 3.1—as a practice that, in expectation, improves veritistic value. It seems we have gone full circle. Starting from the need for single, unified and fruitful listening practice, we have been drawn to Bayesian updating, which in its effectiveness crucially depends on the use of accurate likelihood values. From there, we have looked at a variety of Bayesian source-reliability models whose DAGs encode, and thus allow calculation of this likelihood ratio. Then, upon finding that some of these models tend to produce trusting (> 1) and others distrusting (< 1) results, we started to consider hand-picking them for application in specific reporting contexts, directly undermining the original aim of finding a unified listening practice. I tend to agree with Goldman that this is not the right way to head down. Instead, we should focus our future attention on the overall most adequate and 'rational' among these source-reliability models. As I have argued already and will continue to do in the next chapter, this would be the continuous SD-model.

Before jumping to the next section on local versus global updating, let me quickly address a different route for fine-tuning, above and beyond model choice. As I already mentioned in Section 3.7, a lot could potentially be achieved by deliberately choosing an agent's prior source-reliability credences. We might be especially interested, for example, in the performance of Bayesian agents that are initialized as relatively trusting, or as suspecting anti-reliability, if paired up with advisors that are quite reliable or anti-reliable, respectively. Or, for an example of a different application, taken from Vallinder and Olsson (2014, p. 2002 f.), assume an advisor is reliable, but not extremely so, and a recipient correctly identified them as better than chance. Then, it might be beneficial to this recipient to be overly trusting in their advisor's reliability or to simply be extremely steadfast in their original assessment. Such cases might otherwise run the risk of—due to an 'unlucky' first few pieces of advice—leading to the advisor being incorrectly classified as anti-reliable, which overconfidence or stead-

fastness can stop from happening.[18] An advisor that is reliable to 55%, for example, may start out giving incorrect reports initially and be branded anti-reliable. However, due to the familiar limitations of the subjective perspective, this softer type of fine-tuning is rather unhelpful, because unless the question of advisor reliability is already settled, we can neither properly apply the overconfidence-fix nor hand-pick credence distributions.

4.4 The capability for global updating

There are two different ways a Bayesian source-reliability model may update on repeatedly receiving evidence from a source about a single hypothesis (Merdes, Sydow, and Hahn 2021, p. S5785 ff.): during local updating, the model is applied consecutively to the individual pieces of evidence as they come in, using the posterior of the n^{th} update as the priors for the $n + 1^{th}$. During global updating, the DAG is instead extended to allow simultaneous updates on all pieces of evidence in a single calculation. See Figure 13 for the DAGs required for global updating on just two reports from a single advisor.

However, if reliability is a specific advisor's chance to correctly assess any question in a specific domain of inquiry, why would one even want to update repeatedly in either of these manners? I suggest we interpret this application as a 'soft reset' so to speak: the advisor reconsiders the question or re-evaluates their answer using the same stochastic process as during the formation of their initial advice. Any source with a reliability above 0.5 then still gives testimony that is effectively independent and better than chance, and thus remains helpful continuously. As Condorcet's jury theorem (Dietrich and Spiekermann 2022) tells us, the higher the reliability, and the more of these reports one receives, the closer to one the probability that the majority of them are correct about the hypothesis. Take again the gargoyle example and the question of whether you should enter through the left door, and assume the gargoyle replies affirmatively. If asked again, a perfectly reliable or perfectly anti-reliable gargoyle will simply repeat their original answer. Any other gargoyle might, upon reconsideration, come to respond with a different answer, as their advice is correctly modelled as being generated stochastically.

18. This is somewhat reminiscent of the so-called *Zollman effect* (Zollman 2007, Zollman 2010): here, Bayesian agents may incorrectly classify a theory as worse than it is due to initial bad luck in their inquiry, and subsequently they may fail to ever offer it proper uptake again. Decreasing the amount of communication can help to counteract this risk just like overconfidence does in this source-reliability example, because it allows (some of) the Bayesian reasoners' credences to remain on the 'correct side' of the issue under consideration for just long enough that the initial bad luck is counterbalanced by incoming experiment results or reports.

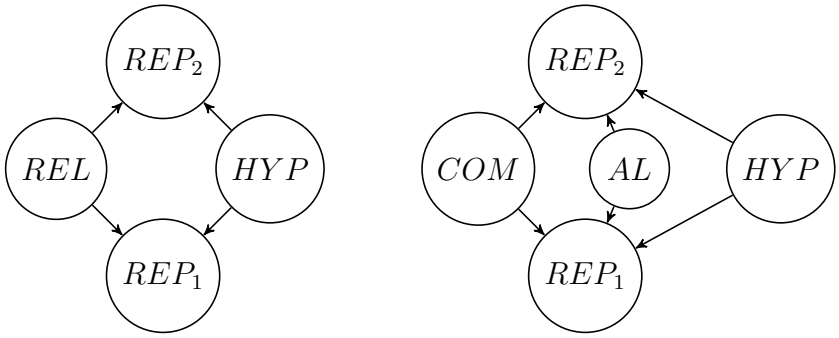

Figure 13: The extensions of the two source-reliability DAGs allow for global updating on two reports from a single advisor concerning a single hypothesis. The DAG on the left (see Merdes, Sydow, and Hahn 2021, p. S5786 and Bovens and Hartmann 2004, p. 76) captures source-reliability in a single node, while the one on the right disentangles competency and alignment.

Repeated local updating on reports about a single hypothesis is a clear divergence from perfect rationality for our Bayesian agents: it forces them to deal with individual pieces of advice at a time, without affording them a 'working memory' of previous testimony above and beyond their current source node credences. This lack of memory makes agents unable to properly handle inconsistent sets of advice unless they are absolutely certain about the truth of the hypothesis (see Section 5.2) and causes "systematic mis-weighting of evidence and order effects" (Ibid., p. S5794, see also Section 5.1). However, even global updating is not without its flaws:

> "[T]he global perspective implies that the focal agent effectively uses the reliability prior for their belief update on all reports (as any sequential outcome is equivalent to updating on all the evidence at once). So even though the global model provides a posterior for REL after obtaining all reports, the fact that the same prior uniquely determines all reports conflicts with the idea that the focal agent is actually dynamically adjusting source reliability after each report in order to adequately evaluate future reports; the agent calculates a reliability posterior, but that reliability posterior does not have any direct impact on the subsequent belief updates. Any 'learning' of reliability in the global model is arguably epiphenomenal with respect to credence in the hypothesis itself." (Ibid., p. S5786)

Treating the estimation source-reliability as a mere after-effect that does not impact an agent's veritistic value with respect to the hypothesis in the slightest means that global updating "arguably fails to adequately address the problem it is trying to solve" (Ibid., p. S5794).

> "Furthermore, 'local' updating seems largely unavoidable for actual, real-world agents: it seems impossible to know in advance what kind of future evidence

> a source may eventually come to report on an issue. The only way an agent could deal with this lack of foresight is to remember all past reports and, after each report received, retrospectively form the appropriate global model and recompute using the initial priors for both reliability and hypothesis. This not only seems unrealistic in practice, it also, once again, renders the reliability revision process itself entirely moot." (Ibid., p. S5787)

What they say here about real-world agents also holds for computationally implemented inhabitants of a simulation. In addition, as soon as a simulation, like the ones discussed in the next section, contains multiple agents taking turns giving advice, the continuity assumed by global updating is broken up:

> "Computational reasons alone already mean that such agents must ignore the network structure: that is, they treat reports from other agents who might themselves be communicating [...] as independent from one another, even though communication creates dependencies. The local application with respect to multiple reports from a single source is in keeping with that limitation." (Ibid., p. S5773)

Taken together, these issues lead me to select local updating for the model used in the next chapter, partly compromising the rationality of these agents in favour of their computational applicability.[19] However, it is crucial to recognize that global updating serves to highlight a cognitive limitation in all Boolean versions of our four source-reliability models that local updating instead obscures. Continuously defined source nodes are perfectly compatible with inconsistent sets of reports in the global updating sense, as for instance 10% negative reports and 90% positive reports are compatible with a whole range of advisor possibilities, whereas Boolean models are not:[20] Thus, Boolean source-reliability models should be regarded highly sceptically, even if we never end up applying them globally. Simulations might necessarily obscure

19. While it would still invite order effects, it might be interesting to explore a mix between global and local updating in future work: this would mean having agents globally update the source-reliability every time a new report comes in, and subsequently perform local updating of the hypothesis using their current reliability estimate.

20. At this point it is possible to provide an interpretation that favours Boolean BH and AL-models, albeit a mostly uninteresting one: if an advisor changing their mind—or diverging from previous reports—about a single hypothesis were reason enough to disregard them as inconsistent, then these models could be said to correctly adjust source-reliability during the global updating procedure. This in turn stands in tension with cases of local model application where advisors are seemingly punished for too much consistency, like when an AL-agent initialized with $p(HYP) = 0.9, p(AL) = 0.6, p(COM) = 0.8$ (or the lowest equivalent α, β-values) repeatedly receives negative reports: despite starting out trusting an advisor who never contradicts themselves the AL-agents will decrease their $p(HYP)$ once, only to then increase it as their now newly distrusted advisor repeats their initial assessment. In this interpretation, it seems that if it were rational to anti-update starting after the second report, then it must have been irrational to update positively at the beginning.

the erroneous nature of these models because they force these Bayesian agents to be amnesiacs, but we should still strive to employ only agents that theoretically *could* stand their ground in a global updating procedure.[21]

When using the Boolean BH-model to update globally on an inconsistent set of reports, an agent must conclude with certainty that the source is a randomizer. Whether the hypothesis is true or false, perfectly reliable advisors will only report their objective truth value. However, inconsistent sets of reports will also contain reports to the contrary. Because according to the Boolean BH-model, an advisor cannot both be truth-tracking and randomizing at the same time, the only reliability value compatible with the data set is 0.5 ($p(REL) = 0$). Note that this holds for any set of inconsistent reports, no matter its size or the ratio of the contained positive and negative reports. Upon receiving a thousand positive reports about the hypothesis and just a single negative one, Boolean BH-agents will conclude that their advisor might as well be flipping a coin. The Boolean version of the AL-model works in essentially the same way, except that it also entertains the possibility of perfectly anti-reliable advisors. For them, as long as a set of reports is internally consistent, it is compatible with truth-tracking advisors whether or not its content matches the actual value of the hypothesis. However, as perfectly reliable sources give a positive report iff the hypothesis is true, and perfectly anti-reliable sources iff the hypothesis is false, Boolean AL-agents too will conclude with certainty that their advisor is a perfect randomizer upon spotting any inconsistencies in a set of reports.

The Boolean version of the (originally continuous) OL-agent fares particularly badly when attempting the global updating procedure. Where for the Boolean BH-agent, any set of reports containing both negative and positive reports was logically inconsistent with the assumption of source-reliability, for the Boolean OL-agent such sets are logically inconsistent full stop. For any value of the underlying question, the Boolean OL-agent can only conceive of sources as reliable, meaning consistently giving reports with only that value, or as anti-reliable, meaning consistently giving reports with the inversion of that value. Recall that neither model, in their Boolean formulations, was able to conceptualize an advisor as anything but perfectly reliable, or perfectly anti-reliable. Without the subjective possibility of stochasticity being involved in the formation of an advisor's report, inconsistent sets of advice about a single hypothesis can simply not be accounted for by any combination of source-node values, and the agent's reasoning breaks down. The Boolean SD-model faces effectively

21. While a detailed examination of global updating with continuous source-reliability models is beyond the scope of this thesis, see Appendix-section 7.7 for an example using the continuous OL-agent.

the same difficulties in accounting for varying reports in the case of global updating on the same proposition. Any inconsistencies in the reports received are regarded as highly problematic because they neither fit the assumptions of full reliability nor of full anti-reliability, which together exhaust its conceptual space. One way of looking at the issue is to view all four Boolean models as equally problematic in characterizing the underlying testimonial situation, with the crucial difference that the OL-agent and SD-agent will stumble over their mistake, realizing the inconsistency between the world and their expectations, whereas the BH-agent and AL-agent will instead wrongly categorize the source as a perfect randomizer.

4.5 Reliability in group simulations

As mentioned throughout the thesis, these focal Bayesian models have been applied as a mechanism for belief revision for agents inhabiting social simulations either directly using the simulation software Laputa (introduced in Section 3.4; see e.g., Pallavicini, Hallsson, and Kappel 2021, Hahn, Hansen, and Olsson 2020, Olsson 2020b), or directly inspired by it (see Assaad 2022). In Laputa, "[f]ollowing Goldman, it is assumed that all inquirers focus on answering one and the same question: whether p or not-p." (Olsson 2011, p. 134). In the last section, I provided an interpretation to make repeated updating on a single hypothesis palpable, one which results in a notion of source-reliability that fits our models: advisors, across all of their reports, stick to a single objective chance of correct advice but may reconsider their reports each round. On the surface, Olsson seems to go for something just like this, defining "Trust (= perceived reliability) is modelled as a second order probability: a credence in the reliability of the source." (2020, p. 4480) and characterizing reliability as the spectrum from "systematically biased/ anti-reliable to systematically truth-telling" (Olsson 2020b, p. 213).

However, as I will argue in this section, Laputa's implementation of source-reliability incurs two related problems: First, it is unclear what exactly reliability means in this model world because agents do not actually possess fixed reliability values that could be properly estimated with the help of Bayesian source-reliability models. Secondly, insofar as there is something close enough, namely some notion of source-reliability$\star$ that is subject to change as a simulation runs, expectation-based 'estimation' of source-reliability is no longer causally independent from the actual presence of source-reliability in a population. Instead, the expectation of anti-reliability can become a self-fulfilling prophecy. Taken together, these worries warrant scepticism concerning the idea that any results from these models pertain to populations of rational agents.

Unlike the model which I used to generate the results for the next chapter or those used by Hahn, Merdes, and Sydow in 2018 and 2021, the advice given by agents in Laputa is not simply stochastically generated based on their competency and alignment values. These agents do have properties that are closely related, namely their aptitude for individual inquiry and their thresholds of assertion, but crucially, the influence of these values is causally mediated by their currently held credences. As a result, even just initializing them with non-neutral priors will throw a wrench in our conception of reliability as a stable, stochastic and law-like relationship between the true value of the hypothesis, and the content of a source's advice.

Additionally, as mentioned in the previous section, the interplay between and interactions of multiple agents in group simulation complicates an otherwise simple picture: advisors update their credences based on their individual inquiry (again dependent on their aptitude values) and on the testimony received by other agents (depending on their current source-reliability estimates). As such, the relationship between the hypothesis' truth value and the content of an advisor's next report is ever-changing.

And of course, whatever an advisor's reliability is ultimately taken to mean, it should remain separable from the question of whether or not they actually happen to be correct about an issue given they already made up their mind. Everyone but perfectly anti-reliable sources can at least sometimes be correct about an issue in the relevant domain of inquiry, and this is especially true for agents that base their testimony partly on their (randomly initialized) prior credences. As such, we ought to be weary of defining an advisor's reliability in relation to the currently held credences.

Previously, idealizations and modelling choices have already compromised even the best focal agents' abilities to figure out objective chances in three main ways: local updating has them misweigh evidence and ignore all but the current report at each point in time; the fact that they perform purely expectation-based updating, rather than mixing in some outcome-based reliability estimation, made them vulnerable to the 'garbage in, garbage out' problem; and only receiving simple Boolean pieces of advice made it impossible for them to effectively distinguish between the first and third, or second and fourth quadrant of the advisor spectrum.[22]

Now, however, an altogether new problem arises, namely the actual lack of an

22. When updating on the experiment results produced by their individual inquiry, Laputa agents use the OL-model to perform expectation-based updating on their own reliability, meaning they do not ultimately have access to purely outcome-based ways of assessing source-reliability. However, at least in updating their self-trust, they are actually attempting to estimate a stable reliability chance. A sufficiently low amount of self-trust may still lead to a vicious circle in individual inquiry (Angere 2010, p. 9), but that will at least not change the underlying aptitude.

underlying, instantiation of unchanging reliability. How is an OL-agent who learns the objective truth about all values in a given Laputa simulation (or alternatively, a Boolean AL-agent in Assaad's simulation) to apply the Principal Principle? For the reasons outlined above, there is no stable objective chance to which they could conform their reliability credence: their assumption of a stable, stochastic relationship between the truth of the central hypothesis and the testimony of the report is undermined by interactions between advisors, by the impact of their individual inquiry, and by the presence of credences that play a central role in the formation of their reports.

One might argue that the recipients are instead simply attempting to track the advisor's current propensity to be correct on the next report, in line with what Angere writes (2010, p. 6), namely that the probability that an advisor that gives a positive report about a true hypothesis with probability r should be assigned a mean reliability credence r—and assuming source-symmetry (see Section 5.3), this value should be the same for a negative report given a false hypothesis. Alternatively still, one might aim to conform subjective credences to the overall frequency of correct reports over the course of the given simulation. However, neither of these interpretations of reliability is completely satisfying, as they are potentially subject to constant change.

As an analogy, imagine you are tasked to figure out the probability with which a coin will land heads by observing an experimenter flip it repeatedly in front of you. However, unbeknownst to you, as the experiment continues they regularly perform sleights of hand to exchange the coin. In doing so, they also change the current objective chance with which the next coin toss will result in heads. As you collect data for your Bernoulli trial, you might well adapt to such a change in objective chance eventually, even if you might lag behind in expectation. However, if someone was to inform you of the true nature of this deceptive game partway through, you would have to do more than simply conform your credence in 'the next coin-flip will land head' to the chances embedded in the currently employed coin. Because estimating an unchanging, objective chance is simply the wrong task for this setup, you would instead have to change your conception of the situation altogether. If you are trying to estimate chances currently embedded in the coin, aware of the fact that that might change, one continuous Bernoulli trial is not the best way to estimate, and an accurate subjective perspective entertains this possibility: assume you observe a million tossed heads in a row, followed by a million tails. Instead of arriving at a steadfast belief (with beta distribution given by $\alpha = \beta = 1000000$) in the fairness of 'the' coin, you might instead simply guess that a coin-switch happened at the one million mark and discard the first half of the coin-toss results. Of course, technically Bernoulli trials should increase veritistic values at every step, even after the coin-switch decreases it

from almost 1 to just above 0.

Similarly, the Bayesian source-reliability models I introduced in the third chapter are effectively trying to estimate stable reliability values. Even to the extent to which Laputa simulations might show that the continuous OL-agents therein may be adequate at repeatedly re-estimating the ever-changing source-reliability[23] of their many advisors, these results would still gloss over this central misunderstanding of the nature of reliability: unable to think outside the box and start the estimation process anew, they will stick to their ongoing process.

Olsson (2013) concludes that polarization in Laputa is evidence that communities of individually rational agents may polarize. He later defends this view against Pallavicini, Hallsson, and Kappel (2021), who argue that this very observation of polarization should instead be regarded as evidence against the rationality of continuous OL-agents as they are embedded in Laputa: in his 2020 article, he responds that "what drives polarization [...] is expectation-based updating, together with a modelling of trust that recognizes the possibility that the source is biased, that is, gives systematically false information." (p. 212). The causal explanation of polarization in Laputa he gives is that if two agents hear one and the same source state that a binary question is to be answered affirmatively, but one of these agents has a high prior in this question and in the reliability of the source, and the other a low prior in both, then they will both become more extreme in their respective beliefs, and further diverge in their beliefs (see p. 218)[24]. Similarly, Assaad argues that expectation-based estimation of source reliability can lead two agents from mere 'surface level disagreements' into 'deep disagreements', i.e., starting only from sufficiently opposing priors about an issue, agents can become locked into mutual distrust and their disagreement may subsequently become irresolvable. By showing that even groups of Boolean BH-agents, which do not 'recognize the possibility of anti-reliability' may polarize if their priors are sufficiently distinct, he also makes the point that expectation-based reliability estimation may cause polarization even in its rather trusting form.

As a result of these dynamics, even if one accepts the notion of source-reliability in Laputa as reliability$\star$, subject to change—whether it is interpreted as current propensities or as overall frequencies—a second problem arises: through the role of advisors' credences in the formation of their reports, their initial priors about the hypothesis and their own expectation of anti-reliability determine their reliability to an unduly

23. Section 5.2 showcases some data on the performance of continuous focal agents when estimating changing reliabilities.

24. Similar points are also made by Henderson and Gebharter in their 2021 paper on belief divergence.

amount. Laputa runs the risk of expectation-based reliability-estimation becoming a self-fulfilling prophecy: even agents with solid aptitude scores and assertion thresholds may be initialized as 'anti-reliable' qua stumbling into incorrect priors and develop or escalate wrong beliefs as a result of inaccurate expectation-based updating on the testimony of others. Thus, framing this as a mechanism for individually rational agents to 'recognise' and avoid anti-reliable sources, as Olsson tends to do, is missing the point: expectation-based estimation of source-reliability may serve to create the very types of agents in a population that it is meant to help avoid. Instead of source-reliability estimation being an epistemic tool that can be separated from and subsequently applied to source-reliability as a property of agents in Laputa, the two are causally intertwined.

Before continuing to the chapter on computational results, I want to briefly mention three ideas on the issue of group simulations with source-reliability models.

First, even if we decide to stick to Laputa mostly as is, I recommend we populate it with agents that reason using the continuous SD-model and properly disentangle misalignment (the threshold of assertion) and competency (a complicated mix of prior veritistic value, aptitude at individual inquiry and the source's own trustworthiness accuracy). This could help with fine-tuning priors to reporting contexts, which may be of interest to at least some research questions: for example, when initializing a population of agents with priors normally distributed around 0.5, with a high mean aptitude and a high mean threshold of assertion, using the continuous SD-model can simplify the choice of prior credences and thus minimize the effects of self-fulfilling expectation of (anti-)reliability.

Secondly, there does exist at least one way of doing 'group simulations' with source-reliability models while simultaneously avoiding the above complications: sticking to the original interpretation of reliability as a stable property of advisors, it might be interesting to write simulation software in which only a single agent reasons using a Bayesian source-reliability model and thus entertains subjective degrees of belief, whereas everyone else is modelled simply as a source with unchanging reliability. While inapplicable to questions about polarizations or filter bubbles, this may still produce interesting results about the performance of source-reliability models themselves.

And thirdly, one might also consider exploring group simulations that focus on more than one hypothesis from one domain of inquiry, such that each recipient only takes a single report from each advisor about each hypothesis, in order better reflect the original idea of competency and interest alignment—and subsequently reliability—as domain-specific properties of advisors.[25] Figure 14 shows the DAG extensions required

25. The epistemic value of such simulations may hinge on the assumption of natural kinds of domains

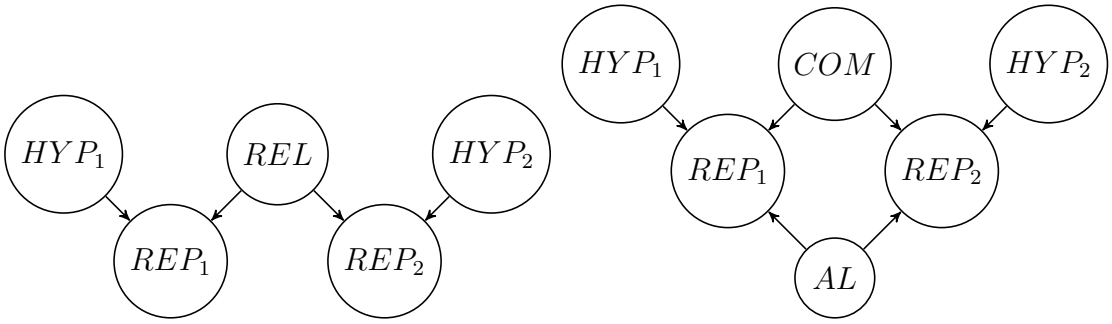

Figure 14: Extended source-reliability DAGs that allow a focal agent to globally update on reports received by a single advisor, and concerning different hypothesis from a shared domain of inquiry. For the DAG featuring a single reliability node (left), see Bovens and Hartmann 2004, for the one disentangling competency and alignment (right), see Assaad 2022.

to perform global updating of source-reliability credence in such setups. Of course, as detailed in the last section, local updating is the more cogent choice for group simulations, due to the alternating impacts from multiple advisors and individual inquiry on the recipients' hypothesis credence. To locally update on the testimony of advisor n about hypothesis k in this manner, the recipient instead plugs in their current (prior or posterior) credence in n's reliability, together with their current (prior or posterior) credence in the truth of k, into the baseline DAG of their respective source-reliability model.

5 Model evaluation: computational results

In terms of Goldman's (1999, p. 104) four stages of the testimony-related activity, my model proceeds as follows: *Discovery* is handled by a stochastic process, during which an advisor assesses the truth value of φ correctly with a probability equal to their competence κ. Next, *production and transmission of messages* is handled via a report $A(\psi)$, which is either equal to the assessment of φ (with probability α) or flipped (with probability $1-\alpha$). This is guaranteed to happen exactly once each round, irrespective of an advisor's confidence or willingness to speak.[26] *Message reception* is rather trivial in the model, with the recipient receiving their advisor's testimony exactly as given. Finally, *message acceptance* is determined—depending on the recipient in question—by the focal Bayesian models introduced in Chapter 3. See Appendix-section 7.8 for

of inquiry across which multiple advisors share equally stable degrees of competency and interest alignment. I do, however, take this to be at least somewhat plausible.

26. As such, the model sidesteps questions regarding 'norms of assertion', such as that one ought to assert the truth of a proposition if one *knows it to be true*, or only if one *has a reasonable belief that it is true*. See Lackey 2007 for a discussion of this issue.

more details about my implementation and for an overview of its user interface.

In order, this chapter features results from this model showcasing how order effects arise from the local application of the focal models, how well they perform source-reliability estimation when they are initialized with perfect veritistic value, how altering the randomization parameter offers up new applications for the AL and BH models, and how the step from Boolean to continuous nodes impacts the steadfastness of agents' reliability estimates.

5.1 Order-dependence of local updating

When discussing the merits of global and local updating in Section 4.4, I mentioned that local applications of source-reliability models lead to misweighing of the evidence depending on the order in which reports are received: after each incoming testimony, the recipient updates their assessment of source-reliability, which will impact their future updates concerning the hypothesis, and, when using continuous models, the future steadfastness of their source-reliability estimates.

For order dependence exploration for the continuous OL-model, see Hahn, Merdes, and Sydow 2018 (p. 667):[27] here the authors compare the agents' posterior degrees of belief in source-reliability and in the hypothesis after feeding them 15 different combinations of two positive and four negative reports. Despite each agent starting with the exact same priors, they observe that "the variability in posteriors is sizeable".

To provide a more intuitive understanding of the development of such order effects for all eight focal agents, Figure 15 contrasts two example runs with a similar setup. Each agent was initialized as somewhat trusting and received either four negative reports followed by two positive ones, or vice versa. I specifically chose a setup for which none (but the Boolean OL-agent) would ever come to assume anti-reliability, in order to ensure generally uniform behaviour across the different agents.

Due to its crucial flaw described in Section 3.4, the Boolean OL-agent avoids displaying any order effects by getting stuck at $p(HYP) = p(REL) = 0.5$ in both runs. The continuous OL-agent, however, does end the first run with a significantly higher credence in the truth of the hypothesis (0.67 to 0.53), whereas their reliability credence remains similar (0.50 to 0.54): in the first run, the trusting agent gives considerable uptake to the two initial, positive reports. In the second run, they are instead confronted with unexpected, negative reports which chip away at $p(REL)$. By

27. After attempting to recreate the results shown in Panel A of Figure 2 in their paper, I am left to assume an accidental mix-up in the order of their graphs.

the time they receive the two positive reports in the second run, they are already much less willing to increase $p(HYP)$, leading to the discrepancy.

The Boolean BH-agent ends the first run with $p(HYP) \approx 0.80$ and only $p(REL) \approx 0.01$, and the second with only $p(HYP) \approx 0.55$, but $p(REL) \approx 0.11$. After initially becoming quite convinced of the hypothesis in the first run, suddenly receiving unexpected, negative reports leads them to drop their reliability estimate far enough that they barely decrease their $p(HYP)$ for the last two negative reports. The continuous BH-agent reacts similarly, though due to its higher steadfastness concerning $p(REL)$, the order effects are mostly averted: $p(HYP) \approx 0.57, p(REL) \approx 0.23$ (reliability expectation of 62%) to $p(HYP) \approx 0.49, p(REL) \approx 0.26$ (reliability expectation of 64%).

Coming next to agents whose models separate source-reliability into competency and alignment nodes, the values I chose for these two runs highlight another aspect: similarly to how the Boolean OL-agent struggles as a result of the symmetric nature of its source-reliability model, these agents can incur difficulty when attempting to disentangle certain credences: if they ever take on the exact same value, the Boolean AL-agent cannot pull $p(HYP), p(AL)$ apart, the continuous SD-agent cannot disentangle competency from alignment, and the Boolean SD-agent will even keep all three of their credences constantly equal to one another.

As a result, the Boolean AL-agent too behaves eerily similar to the Boolean BH-agent, except for their competency prior of $p(COM) = 0.6$ being higher that the BH-agent's reliability prior of $p(REL) = 0.36$. This difference, of course, is the result of initializing these two agents as equally trusting overall, with the AL-agent assuming an imperfect degree of interest alignment. For the continuous AL-agent, whose alignment credence distribution is considerably more steadfast than their credence in the hypothesis, the initial equality between $p(HYP)$ and $M(r)$ is quickly broken up and the behaviour resembles that of the continuous BH-agent: they end the first run with $p(HYP) \approx 0.56$ and an expectation of reliability equal of about 0.63, and the second run with $p(HYP) \approx 0.49$ and estimating their advisor's reliability to remain around 63%.

And finally, the Boolean version of the SD-agent, with three symmetrical source nodes all initialized with a prior of 0.8 cannot disentangle them in any way, shape or form: it ends with $p(COM) = p(AL) = p(HYP) \approx 0.57$ in the first run, and approximately 60% in the second. The continuous version manages to separate their $p(HYP)$ from the means of their identical credence density distributions for both competency and alignment, which encode a source-reliability expectation of 60% at the end of the first run, and of 0.62 at the end of the second run. Meanwhile, their $p(HYP)$ drops from 0.57 to 0.50.

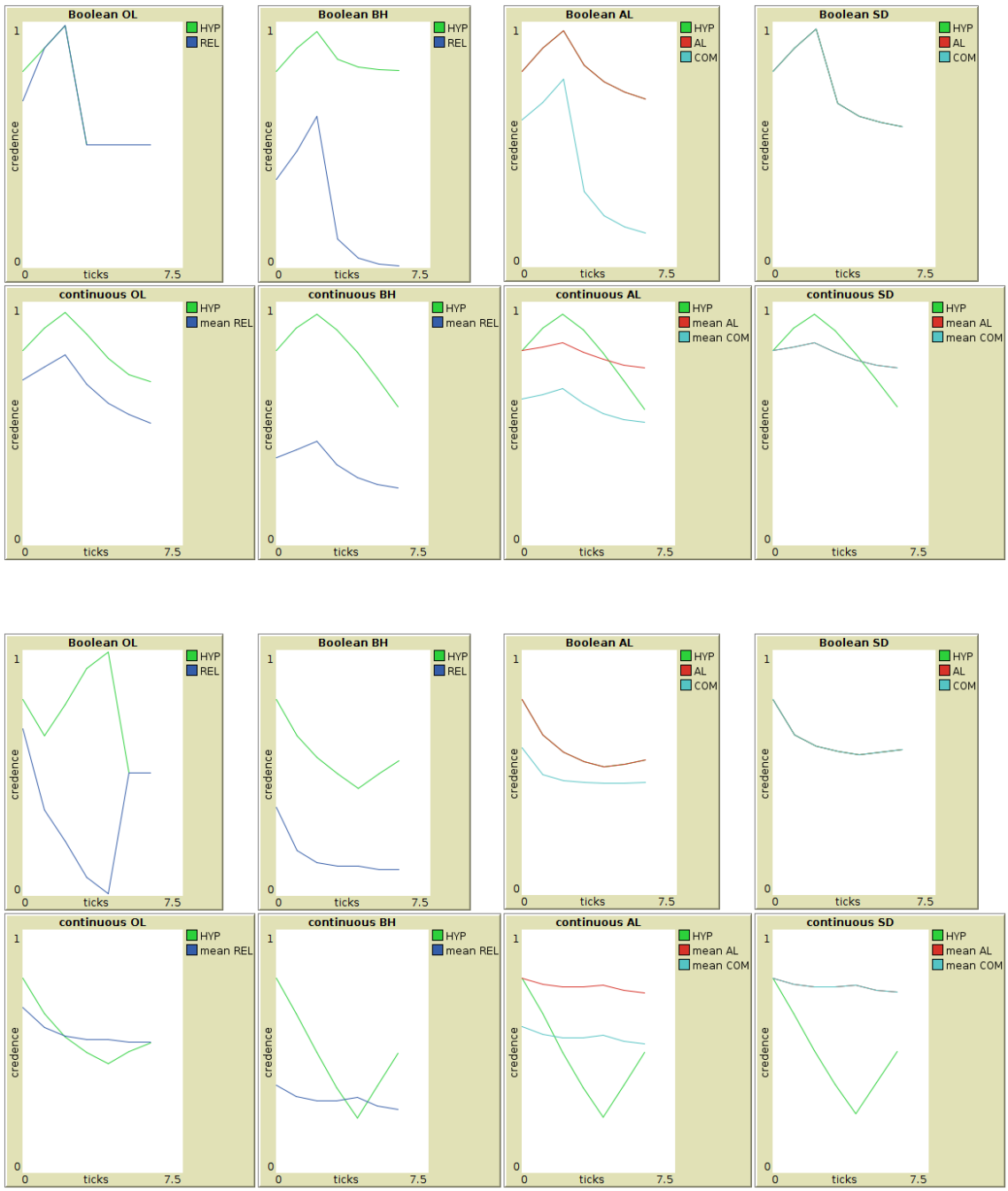

Figure 15: Example of order dependence of the local updating procedure: all eight focal models were initialized with the usual trusting priors (see Table 1), assigned a (mean) credence of 0.8 in the hypothesis and received a total of six reports. The agents above received two positive reports followed by four negative ones, the agents below received these reports in reverse order. Except for the flawed Boolean version of the *OL*-agent, every focal model posterior shows dependence on the order of the received reports.

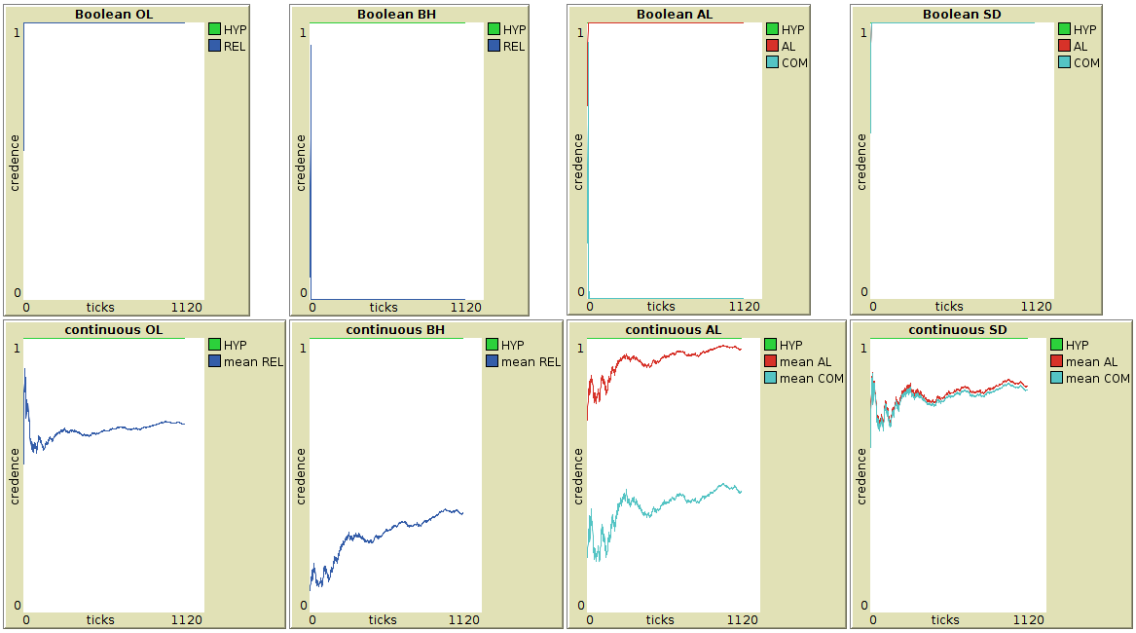

Figure 16: Example model run with every agent simply knowing that the hypothesis is true. Focal agents were initialized as sightly trusting, with priors encoding an expectation of advisor competency of 0.6, and degree of interest alignment of 0.7. In fact, however, the objective $\alpha = \kappa = 0.8$, resulting in a 0.68 propensity to give correct reports and an actual frequency of 68.8% in this run specifically. The model was run for 1000 ticks and the resolution of the unit interval was set to 1000.

5.2 Gold in, gold out?

In Section 3.1 I discussed the limits of expectation-based updating in terms of the 'garbage in, garbage out' problem. Receiving binary reports from a single source in an outcome-free vacuum, priors are everything for these source-reliability models, and hence, computationally exploring their behaviour for the entire parameter space is not a fruitful endeavour. Having understood their updating procedures conceptually, how they will behave for positive or negative reports for any combination of source-node priors is within reach intuitively.

However, there is a flip-side to 'garbage in, garbage out' that I want to address in this section: how well do these models fare in estimating source-reliability if they happen to 'know' the true value of the hypothesis, that is, when they entertain accurate credences $p(HYP) = 1$?[28] Or in other words, what happens if the Bayesian agents are fed 'golden priors'? To isolate the expectation-based estimation of source-reliability performed by these focal Bayesian models, Figures 16 and 17 each show an example

28. Generally speaking, Bayesian agents ought to avoid assigning the extreme values variables values $0, 1$, as they are incapable of reasoning themselves out of these corners using conditionalization on incoming evidence. In this case, however, this property works in our favour.

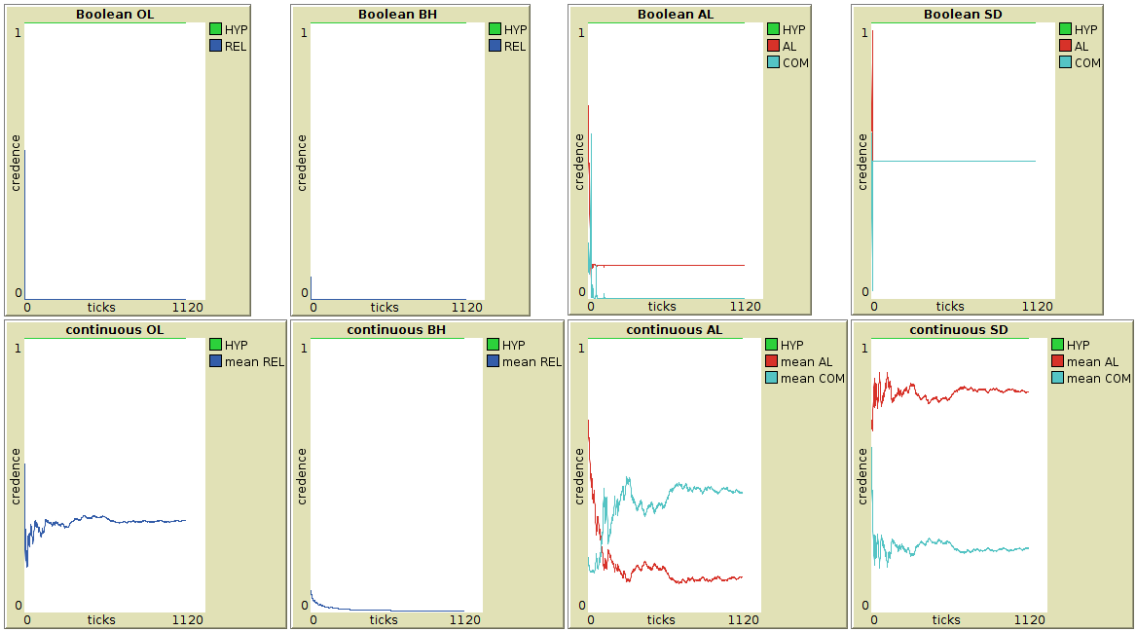

Figure 17: Contrast to the run in Figure 16: everything was kept identical, except that an $\alpha = 0.2$ and a $\kappa = 0.8$ now combine into an anti-reliable advisor who gave correct reports only 33.4% of the time (propensity of 32%).

run in which they are confronted with 1000 consecutive reports from a single source. I chose this high number of rounds to highlight the approximation of the actual frequencies in expectation, which I take to be visually more helpful than displaying the mean of many runs with 10 pieces of advice each.

As Figure 16 shows, for a setup like this, the Boolean source-reliability models could be said to 'take in gold and put out garbage'. They either quickly settle on incorrect reliability estimates and stick to them, or in the case of the Boolean OL and SD-agents, even break down entirely, as they cannot update on negative reports when entertaining $p(HYP) = p(REL) = 1$ (or $p(HYP) = p(COM) = p(AL) = 1$). The continuous agents, on the other hand, faired much better: despite being initialized with incorrect source-reliability priors, they managed to turn their golden hypothesis priors into quite accurate estimations of source-reliability. For the OL-agent and BH-agent, which stay agnostic about the exact composition of source-reliability, their pure reliability estimates are theoretically compatible with infinitely many combinations of α, κ values, and do not even differentiate between the first or third quadrant. For the SD-agent and AL-agent, the limited epistemic environment provided by my computational model meant that the best they can do is settle onto one combination of mean competency and alignment credences that approximate the observed frequency of agreeable testimonies.

The continuous OL-agent ended their run with a mean reliability credence of

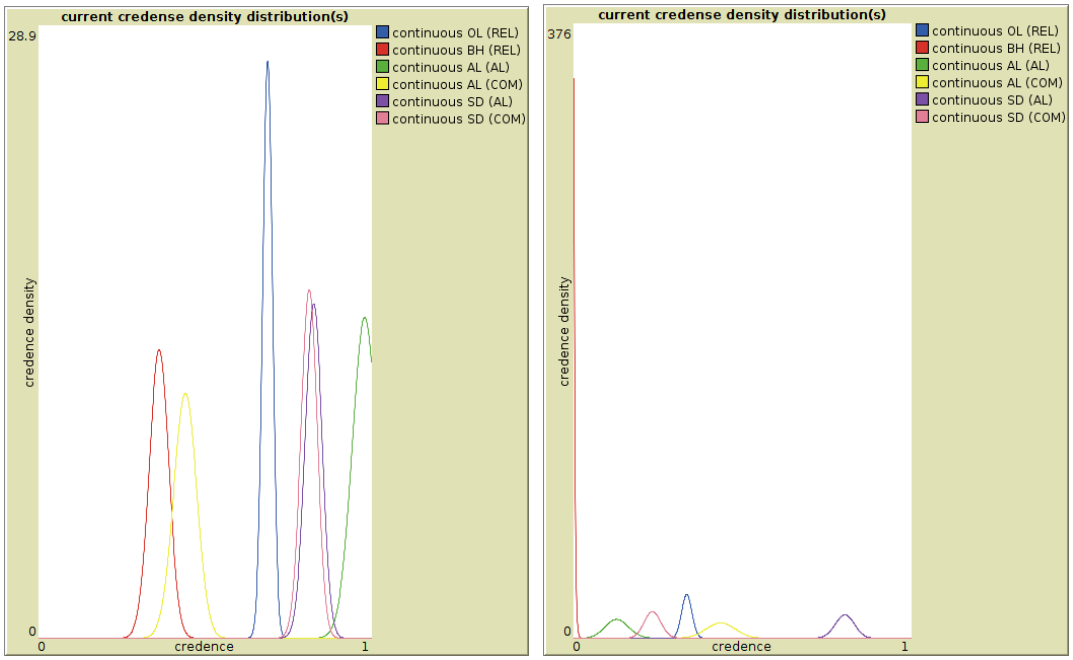

Figure 18: Final credence density distributions of the continuous source-reliability models for the two example runs contrasted in this section. Over their 1000 updating rounds, these agents became quite steadfast in their beliefs, as encoded in the steepness and low variance of the curves.

roughly 0.69 (0.68 is accurate to the underlying propensity, 0.69 to the actual frequency), the continuous BH-agent with 0.36 (either 0.36 or 0.38 is accurate), the continuous AL-agent with a mean alignment credence of 0.96 and mean competency credence of 0.44, which taken together encode an expectation 70% source-reliability. The continuous SD-agent arrived at mean credences of 0.83 in the advisor's alignment, and of 0.81 in their competency, which together expresses a reliability credence of about 0.70.

Figure 17 gives an overview of the results of a similar run with the exact same initialization, except that this time, the objective competency, alignment and reliability values were $\kappa = 0.8, \alpha = 0.2, \alpha\kappa + \overline{\alpha}\overline{\kappa} = 0.32$. The mean reliability credence of the OL-agent ended up at around 0.33 (0.32 would be accurate to the propensity, 0.33 to the actual frequency in this run), the AL-agents mean alignment credence at 0.13 and their mean competency credence at about 0.43, encoding a reliability estimate of 0.34, and the SD concluded with a mean alignment credence of 0.80, mean competency credence of 0.23, and a subsequent reliability estimate of $0,34$.

Figure 19 displays the shape of the agents' credence density distributions at the end of both runs side-by-side. The most crucial difference in this run is that as it features an anti-reliable advisor, the continuous BH-model was unable to properly approximate

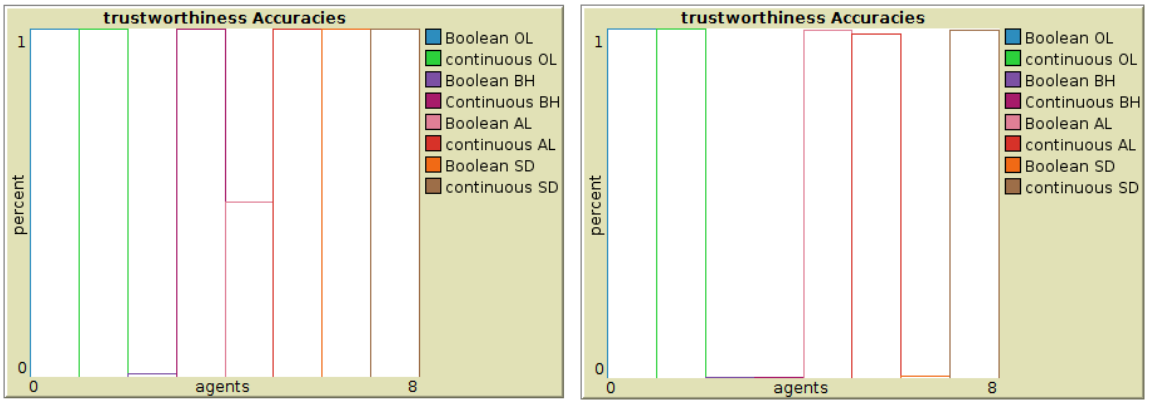

Figure 19: Trustworthiness-accuracy scores for all eight focal agents in the two runs contrasted in this section. As they were initialized as slightly trusting, in the run featuring a reliable source (left), most agents received high scores. The only exceptions are the Boolean BH and AL-agents, which incorrectly classified the source as a randomizer. Compared to the run featuring anti-reliability (right), the biggest difference is that due to its trusting nature, even the continuous BH-agent earned a score of 0.

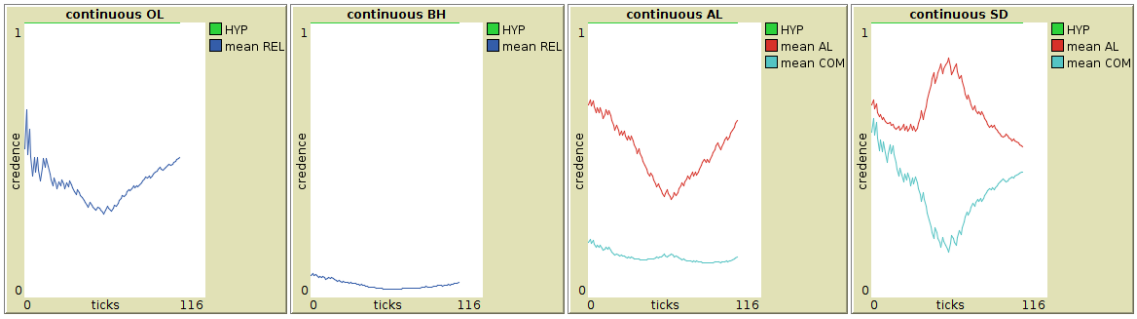

Figure 20: Continuous focal agents with golden hypothesis priors reacting to a change in source reliability: initialized as slightly trusting as listed in Table 2, they first faced 50 reports from a source with a reliability of 0.18, followed by 50 reports from a source with 82% reliability, combining to a total of 51 out of 100 correct reports.

their reliability, but could at best decrease their mean reliability credence to almost 0 (≈ 0.003), the expectation of an almost perfectly randomizing advisor. As Figure 19 therefore shows, is that the continuous BH-agent is the only continuous agent with 0% trustworthiness accuracy in this run.

Lastly, I mentioned in Section 4.5 how, similarly to Bernoulli trials struggling with estimating the objective chance that a coin will land heads if the coin is replaced halfway through the trial, these Bayesian source-reliability models are no perfect fit for changing source-reliabilities. Figure 20 shows how the four continuous versions deal with estimating the reliability of an advisor that starts low (at 0.32) and suddenly jumps high (to 0.64) after 50 rounds. As expected, while the agents do immediately reverse course, the baggage of their previous experience, as encoded in their steadfast

distributions, holds them back considerably. At the end of the simulation, the SD-agent and AL-agent estimate source-reliability at around 50% versus around 52%, the slightly less steadfast OL-agent at 51% and the BH-agent who had struggled to account for the anti-reliability during the first half of the simulation around 53%. This run also serves to highlight the tendency of AL-agents to adjust their alignment credences more quickly than their competency credences.

5.3 The randomization parameter

Some of the focal models introduced in the last chapter, namely (all variations of) the AL-agent and BH-agent feature a so-called randomization parameter β, meant to encode the probability that an advisor gives a positive report when they randomize based on their competency. OL-agents and SD-agents, capable of fully conceptualizing the competency axis, do not feature a randomization parameter: Here, competency-based randomizers always give positive reports with a probability of 50%. Bovens and Hartmann establish that by default, the value of β ought to be $\frac{1}{n}$, with n encoding the number of possible answers to the question at hand, which comes down to a default value of $\frac{1}{2}$ for Boolean questions. However, there are two different reasons one might want to set $\beta \neq 0.5$:[29] First, to accurately model randomizing advisors with certain asymmetric propensities to give positive/ negative reports, and (ii) to model special, simplified cases that restrict the advisor spectrum.[30]

First, adjusting β within the unit interval allows us to account not only for perfect randomizers 'flipping fair coins', but also for those with a tendency akin to a weighted coin. As an example of extreme cases, take 'yay-sayers' who answer every question posed to them affirmatively, or 'nay-sayers' who deny any question posed to them. Such advisors clearly do not pick up on actual features of the world, so at first glance, the best way to represent them is by setting the expectation of reliability to 0.5 and the probability that they give a positive report given they are unreliable to exactly 1 or exactly 0. And while yay/nay-sayers can be safely taken to be exceptionally rare, weaker versions of randomizers with skewed propensities should be expected to occur more frequently.

When plugging $\beta = 1$ into the Boolean BH-model, it becomes a model suited for

29. In any case, however, one should be certain of the appropriate value for the randomization parameter because leaving that open greatly undermines the model's predictive power (see also Osimani and Landes 2020).

30. I restrict my analysis in this section to a single piece of testimony concerning a single hypothesis, a setup for which the global and local updating procedures are equivalent.

sorting out yay-sayers. As Figure 21 shows, upon receiving just a single negative report, the Boolean BH-agent will immediately jump to the conclusion that their advisor is perfectly reliable—after all, otherwise, they would necessarily give a positive report—and consequently trust that the hypothesis must indeed be false. From here on out, receiving a positive report would appear inconsistent to the Boolean BH-agent, as it would require a perfectly reliable advisor to give testimony contradicting their certain belief that the hypothesis is false. Similarly, the continuous BH-agent too immediately concludes the falsity of the hypothesis. However, for them the negative report is not conclusive evidence for perfect reliability, because it is compatible with a whole range of partly reliable sources that just so happen to give accurate testimony in the first round, see Figure 22 for the shape and change in their credence distribution. Receiving a positive report instead will lead the Boolean BH-agent to slightly increase their belief in the hypothesis, and slightly decrease their reliability credence: the report is correctly interpreted as evidence against the only possibility that could have produced a negative result, namely a reliable advisor reporting about a false hypothesis.[31] The continuous BH-model reacts in the same manner, except for its comparatively high steadfastness in their mean reliability credence (see Section 5.4).

As long as their degree of interest alignment exceeds 0.5, AL-agents react essentially in the same way: upon receiving a negative report, which can only be the result of competency, the Boolean AL-agent's competency credence jumps to 1 and the continuous AL-agent increases theirs significantly. Both react rather tranquilly to positive reports, which can easily be explained away as the result of yay-saying. However, because AL-agents are weary of possible deception, they do not immediately become certain that the hypothesis is false upon receiving a negative report. Initializing them with a (mean) alignment credence below 0.5 even flips the direction in which they change $p(HYP)$. For example, a negative report is still ascribed to an advisor's competency, but an AL-agent assuming anti-alignment takes this very fact as evidence for the hypothesis.

The second application of $\beta \neq \frac{1}{2}$ is what I will call *epistemic zoom*: many times in testimonial situations, recipients may not be interested in the entirety of the advisor-spectrum, but rather need to merely distinguish a subset of the conceptual possibilities. As an example that favours the Boolean AL and BH-agents, imagine you are present

31. As Henderson and Gebharter (2021) show, upon receiving a positive report, the Boolean BH-model increases their reliability estimate iff $p(HYP) > \beta$. Effectively, this condition reflects a β threshold below which positive reports are better explained by a reliable source confirming the recipient's belief in the hypothesis, rather than by a randomizer stumbling into giving a confirmatory report.

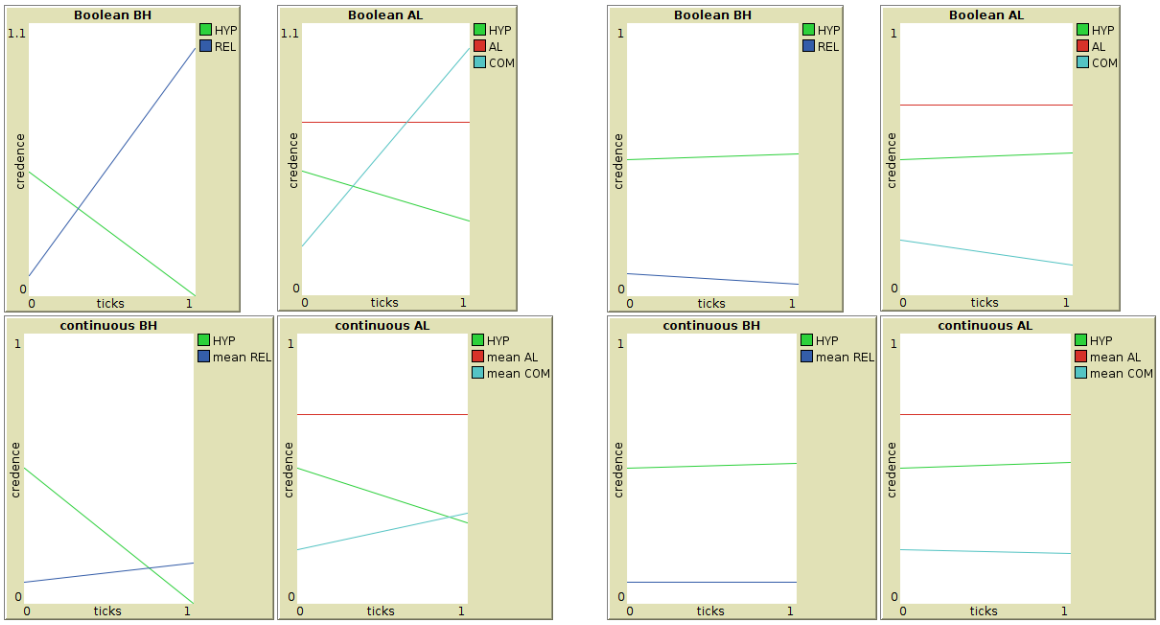

Figure 21: Boolean and continuous BH-agents and AL-agents reacting to a single negative report (left), versus to a single positive report (right). Agents were initialized with $p(HYP) = 0.5$, their randomization parameter set to 1, and as slightly trusting (see Table 2).

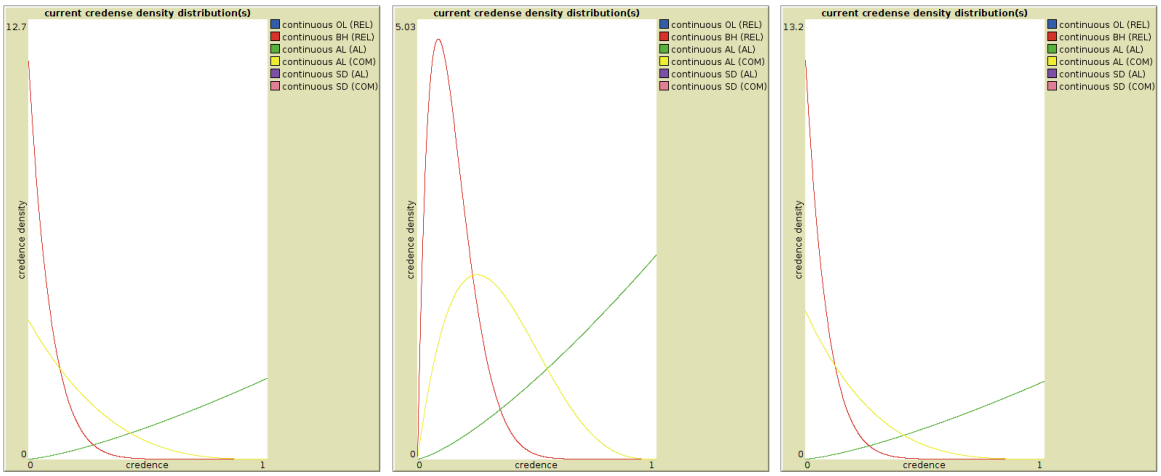

Figure 22: Initial credence density distribution of the continuous BH-agent and AL-agent (left), compared to after they receive a single positive (center) or negative (right) report.

when for the experimental testing of whether a self-proclaimed mind-reader is in truth a charlatan. Assume that an experimenter throws two 10-sided dice in such a way that only they, but neither you nor the subject, can observe the outcome, and asks the alleged mind-reader whether the dice show doublets. In this case, we can assume perfect interest alignment, given that the subject seeks to convince everyone of their capabilities, and as such, we are free to replace the AL-agent with the now equivalent BH-agent's with reliability mapped directly onto the competency axis. For this Boolean question, a true, competence-based randomizer (modelled by $\beta = 0.5$) would be correct exactly half of the time. However, even charlatans should be expected—qua understanding basic probability—to be much better than chance, which we can account for by setting $\beta = 0.1$.[32] Now the BH-agent is no longer concerned with differentiating between advisors with 50% or 100% competency, but rather tasked with figuring out whether the test subject is perfectly reliable (able to observe the outcome of the die-roll by reading your mind), or a someone who is better than chance by understanding die-rolls. For such a setup then, the Boolean BH-agent captures the intuition that hearing the subject claim doublets should raise your credence in their powers by a greater factor than hearing them give a negative report should lower it.

In some sense, then, models without a randomization parameter appear less flexible with respect to possible application scenarios. However, both of these application cases diverge from the objective model sufficiently to be disregarded for the remainder of this thesis.

Firstly, propensities towards positive or negative reports may well be a serious feature of the testimonial situation, but they can already be captured by the objective model because they are encoded in the ratio between an advisor's reliability for the subset of true questions in a domain, and their reliability for the subset of false questions in the same domain. As such, they can be modelled with separate reliabilities for both subsets, which also capture the fact that yay/nay-sayer-propensities do not merely exist for exactly 50% reliability. For example, the incorrect reports given by an advisor with 90% reliability may be evenly split into equal numbers of 'false negatives' and 'false positives', but if they have a bias in either direction, it might be fruitful to disentangle the two subsets altogether. Vallinder and Olsson (2014, p. 1993) and

32. In this example it is easy to see why β would be set to 0.1, as 10 out of the 100 possible outcomes of the dice roll are doublets. However, choosing the correct value for the randomization parameter can be slightly more involved at times. As another example, say the BH-agent is tasked to discern whether someone known to have recently purchased a car bought one of type A. Knowing nothing else, $\beta = \frac{1}{2}$ might be a reasonable assumption. However, if it is common knowledge that car A has a 90% market share, one should model the epistemic baseline expectation of what someone who knows nothing additional about the purchase would answer by increasing β to 0.9.

Angere (2010, p. 6) discuss such ideas under the name of 'source symmetry', the assumption that a source's reliability with respect to φ equals their reliability with respect to $\neg\varphi$. This assumption rules out cases in which an agent is more likely to detect that φ if φ is true, than that $\neg\varphi$ if φ is false, or vice versa. Olsson (2011) classifies updating on the advice of such symmetric sources as a 'truth invariant' epistemic practice. Updating on the testimony of even a partial yay/nay-sayer, who count as asymmetric sources, would then count as a 'truth variant' practice.

Secondly, when changing β to perform epistemic-zoom, the recipient has to give up on the idea that actual randomizing by their advisor would result in a 50% chance of correct advice, and thus temporarily change what is meant by 'unreliable source'. If a recipient is aware of such artificial limits of the advisor space in advance, such special cases are indeed a useful application for the Boolean BH and AL-models. Otherwise, the generally more comprehensive strategy is to let a focal agent with a continuously defined notion of source-reliability figure out the exact value, keeping with the OD-model in being open to all the possibilities. In the magician case, this might for instance serve to distinguish charlatans (more reliable than chance qua understanding dice) and true mind-readers (perfectly reliable) from actual coin-flippers (unreliable), who could not be modelled by the BH-model while one uses it to 'zoom-in'.

5.4 Steadfastness

As argued in Section 3.2, one of the two major upsides to outfitting source-reliability models with continuous nodes for reliability or competency and alignment is that it allows them to store information above and beyond a simple credence: it allows them to differentiate between different amounts of certainty. This section highlights an example of how the presence of continuous source nodes —as well as their initialization—can impact the steadfastness of Bayesian agents with respect to source-reliability.

When computationally exploring their behaviour, Merdes, Sydow, and Hahn initialize their continuous OL-agents as rather undecided: to arrive at the expected reliability value of $M(r) = \frac{2}{3}$, which they take to represent a compromise between healthy scepticism and the assumption that the source is more reliable than not, they chose the lowest possible values of $\alpha = 2, \beta = 1$ (see 2021, p. 5788). A thus initialized agent will, similarly to their Boolean counterparts, quickly change their mind about a source's reliability. This allows easier comparison between the two types of agents, but also hides a potential advantage for the continuous agents: not only may they become more steadfast as time goes on, but if there is sufficient reason to initialize them as trusting we even might want to initialize them as staunchly trusting in the first place,

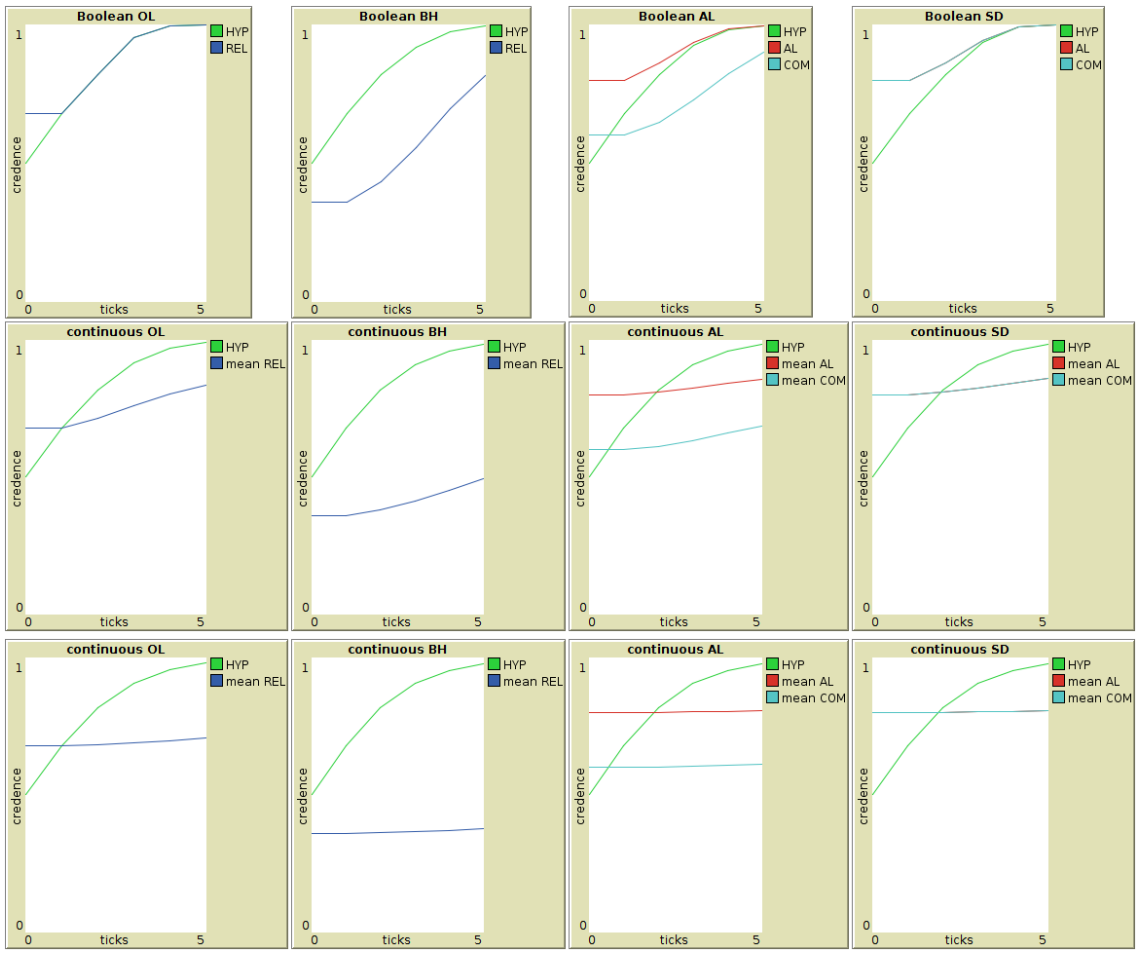

Figure 23: Steadfastness comparison between Boolean agents (top row) and continuous agents (middle row) initialized as relatively trusting with priors described in Table 1, and continuous agents who were initialized with steeper beta-density distributions (bottom row) with the same expected values, but with α, β each multiplied by 10. All agents started with prior $p(HYP) = 0.5$ and received five pieces of confirmatory testimony.

by multiplying their α and β values by the same factor.

Figure 23 compares the steadfastness of three different agents per source-reliability model, as they all start with $p(HYP) = 0.5$ and receive five consecutive pieces of positive reports. Boolean agents are by far the quickest to increase their reliability or competency and alignment credences, and consequently also accelerated the speed at which $p(HYP)$ approaches one. The continuous agents initialized with credence density functions with the highest compatible variance are already considerably more steadfast in their reliability-related belief revision and thus lag slightly behind their Boolean counterparts with respect to their hypothesis credences. And lastly, multiplying the α and β values for the initial credence density functions results in continuous

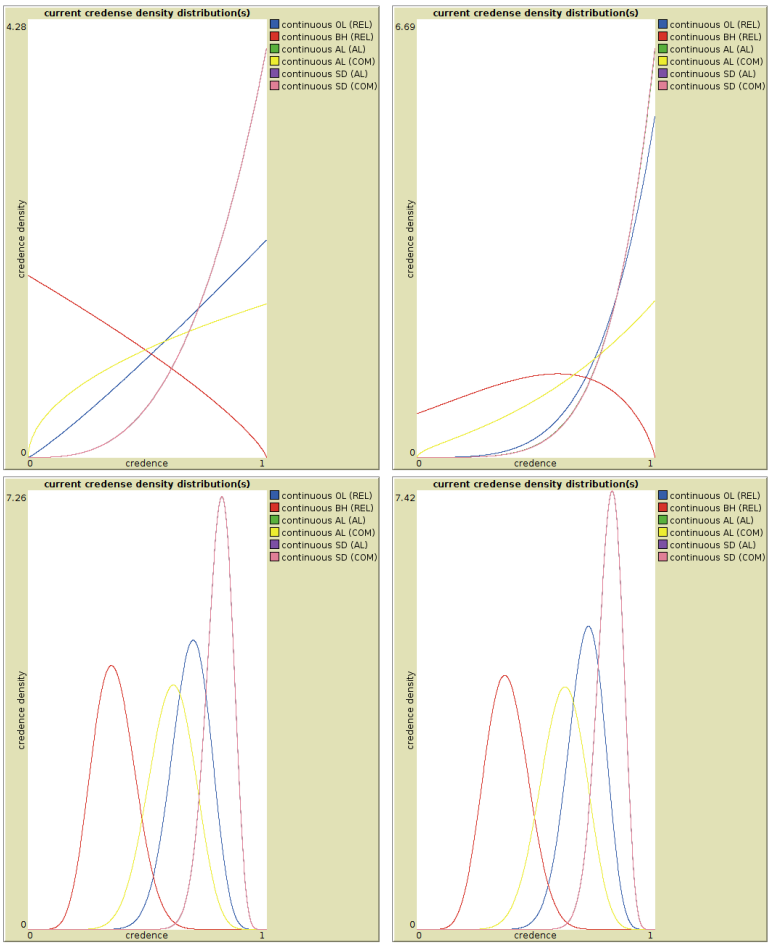

Figure 24: Development of the credence density distributions from their initialization (left) to after updating on five consecutive, positive reports (right), and the contrast between agents initialized as relatively uncertain (above) and relatively certain (below).

agents with extremely steadfast reliability estimates that barely change as the reports are received, and as such, they are also by far the slowest to approach $p(HYP)$.

Figure 24 highlights the differences between the credence density distributions of these two types of continuous agents. When initialized as steadfast, they start out with steep, low variance distributions that barely change in shape or expected value as the agent updates on the five reports. If they are instead initialized as relatively uncertain, their prior distributions are much less steep and undergo noticeable shifts and shape-change throughout the five rounds.

6 Conclusion

This thesis improved on a way of describing the testimonial situation, and more specifically, an advisor's reliability, in terms of domain-specific competency values and a

degree of interest alignment between an advisor and the recipient of their advice. I argued that while on its own, this description of the testimonial situation is limited because outside of the inaccessible 'God's eye' perspective, it does not help the recipient discriminate between the many possible combinations of advisor competency and alignment, it can still be (i) used to conceptually clarify the target phenomenon and (ii) employed in computational modelling.

Drawing from the clarity provided by the objective model, I have then presented Boolean and continuous versions of four Bayesian source-reliability models and categorized each of them in accordance with which sections of the advisor spectrum they can represent. Each of them provides a different account of how one ought to update one's credences in a source's reliability and the truth of a Boolean hypothesis based on (expectations of) incoming testimony.

The extent to which (an application of) these models can be regarded as adequate—above and beyond simply meeting the criteria for 'Bayesian rationality'—to the situation varies along many axes: are they able to disentangle competency and degree of interest alignment by splitting their reliability node into two separate source nodes? Do they entertain the possibility of anti-reliability? Are they capable of conceptualizing whole areas of the advisor spectrum, instead of merely individual points? And finally, can they distinguish between different amounts of information leading to the same expectation of source-reliability?

The continuous SD-model is the only one for which the answer to all of these questions is 'yes', as it combines the best aspects from the three established models: agents that use it can handle inconsistent data sets, fully conceptualize the advisor spectrum by disentangling reliability into fully continuous alignment and competency nodes, and can store information about steadfastness and certainty in the shape of its credence distributions. This is why, whenever we find fruitful applications for such models, I suggest we should use the continuous SD-model, rather than taking on the habit of hand-picking models depending on the listening context we might apply them to.

And yet, there remain severe problems for future projects wishing to employ Bayesian source-reliability models. Computational limitations force us to compromise the 'rationality' of our Bayesian agents: using them locally (through repeated application of the same, static DAG, rather than globally via extending the DAG as necessary) has them misweigh evidence based on the order in which they receive testimony, and require the choice of discrete resolutions to approximate the distributions of second-order credences over otherwise continuous source nodes. Additionally, their expectation-based reasoning suffers from the 'garbage in, garbage out' problem

common to formal reasoning systems, and is therefore not very useful on its own. Even upon settling for local updating with continuous SD-models that combine their expectation-based updating with some form of outcome-based source-reliability estimation, the question of their proper application to group simulations remains unsolved, as I have argued in Section 4.5: the focal agents are attempting to estimate an unchanging reliability value, which the model objectively does not feature, and this mismatch undermines their adequacy to the task. To make matters worse, this results in a possibility of expectation-based reliability estimation itself being a cause of 'anti-reliability' (in the malleable sense) in these models, which pulls the normative scope of their results into question.

Overall, I hope this thesis managed to provide an insightful overview and explanation of Bayesian source-reliability models and to help ease their future applications, while simultaneously motivating a healthy scepticism about concluding the normativity of the behaviour displayed by Bayesian agents whose reasoning is based on them.

Acknowledgements

Thanks to Stephan Hartmann and Hein Duijf for supervising this thesis; to Erik J. Olsson for granting me access to a Linux version of Laputa; to Christoph Merdes for insight into the computational details behind the 2021 paper he co-authored with Ulrike Hahn and Momme von Sydow; and especially to Leon Assaad, Leyla Ade, Korbinian Friedl and Sam Swartzberg for helpful discussions and feedback throughout.

References

Angere, Sofie. 2010. "Knowledge in a social network." *Synthese,* 167–203.

Assaad, Leon. 2022. "Trust in Information Sources: Reconstructing Worrisome Phenomena Using Bayesian Models of Source Reliability."

Baron, Jonathan. 1987. "Second-order probabilities and belief functions." *Theory and Decision* 23 (1): 25–36.

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

Brier, Glenn W, et al. 1950. "Verification of forecasts expressed in terms of probability." *Monthly weather review* 78 (1): 1–3.

Dietrich, Franz, and Kai Spiekermann. 2022. "Jury Theorems." In *The Stanford Encyclopedia of Philosophy,* Summer 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

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

Goldman, Alvin I. 1999. *Knowledge in a social world.* Oxford University Press.

———. 2001a. "Experts: Which ones should you trust?" *Philosophy and phenomenological research* 63 (1): 85–110.

———. 2001b. "Quasi-objective Bayesianism and legal evidence." *Jurimetrics* 42:237.

———. 2018. "Expertise." *Topoi* 37 (1): 3–10.

Hahn, Ulrike, Jens Ulrik Hansen, and Erik J Olsson. 2020. "Truth tracking performance of social networks: How connectivity and clustering can make groups less competent." *Synthese* 197 (4): 1511–1541.

Hahn, Ulrike, Adam JL Harris, and Adam Corner. 2009. "Argument content and argument source: An exploration." *Informal Logic* 29 (4): 337–367.

Hahn, Ulrike, Christoph Merdes, and Momme von Sydow. 2018. "How good is your evidence and how would you know?" *Topics in Cognitive Science* 10 (4): 660–678.

Hahn, Ulrike, and Mike Oaksford. 2007. "The rationality of informal argumentation: a Bayesian approach to reasoning fallacies." *Psychological review* 114 (3): 704.

Hahn, Ulrike, Mike Oaksford, and Adam JL Harris. 2013. "Testimony and argument: A Bayesian perspective." In *Bayesian argumentation,* 15–38. Springer.

Hájek, Alan, and Stephan Hartmann. 2010. "Bayesian epistemology."

Harris, Adam JL, Ulrike Hahn, Jens K Madsen, and Anne S Hsu. 2016. "The appeal to expert opinion: Quantitative support for a Bayesian network approach." *Cognitive Science* 40 (6): 1496–1533.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array programming with NumPy." *Nature* 585 (7825): 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Heinzelmann, Nora, and Stephan Hartmann. 2022. "Deliberation and confidence change." *Synthese* 200 (1): 42.

Henderson, Leah, and Alexander Gebharter. 2021. "The role of source reliability in belief polarisation." *Synthese* 199 (3-4): 10253–10276.

Howson, Colin, and Peter Urbach. 2006. *Scientific reasoning: the Bayesian approach.* Open Court Publishing.

Jarvstad, Andreas, and Ulrike Hahn. 2011. "Source reliability and the conjunction fallacy." *Cognitive Science* 35 (4): 682–711.

Joyce, James M. 2005. "How probabilities reflect evidence." *Philosophical perspectives* 19:153–178.

Lackey, Jennifer. 2007. "Norms of assertion." *Noûs* 41 (4): 594–626.

Lewis, David. 1980. "A subjectivist's guide to objective chance." In *Ifs,* 267–297. Springer.

Merdes, Christoph, Momme von Sydow, and Ulrike Hahn. 2021. "Formal models of source reliability." *Synthese* 198 (23): 5773–5801.

Oaksford, Mike, and Ulrike Hahn. 2012. "Why are we convinced by the ad hominem argument?: Bayesian source reliability and pragma-dialectical discussion rules." In *Bayesian argumentation: The practical side of probability,* 39–58. Springer.

Olsson, Erik J. 2011. "A simulation approach to veritistic social epistemology." *Episteme* 8 (2): 127–143.

———. 2013. "A Bayesian simulation model of group deliberation and polarization." In *Bayesian argumentation,* 113–133. Springer.

———. 2020a. "A diachronic perspective on peer disagreement in veritistic social epistemology." *Synthese* 197 (10): 4475–4493.

———. 2020b. "Why Bayesian agents polarize." In *The Epistemology of Group Disagreement,* 211–229. Routledge.

Osimani, Barbara, and Juergen Landes. 2020. "Varieties of error and varieties of evidence in scientific inference."

Pallavicini, Josefine, Bjørn Hallsson, and Klemens Kappel. 2021. "Polarization in groups of Bayesian agents." *Synthese* 198 (1): 1–55.

Rini, Regina. 2021. "Weaponized Skepticism." *Political Epistemology,* 31.

Shafto, Patrick, Baxter Eaves, Daniel J Navarro, and Andrew Perfors. 2012. "Epistemic trust: Modeling children's reasoning about others' knowledge and intent." *Developmental science* 15 (3): 436–447.

Tversky, Amos, and Daniel Kahneman. 1983. "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment." *Psychological review* 90 (4): 293.

Vallinder, Aron, and Erik J Olsson. 2014. "Trust and the value of overconfidence: A Bayesian perspective on social network communication." *Synthese* 191 (9): 1991–2007.

Van Rossum, Fred L., Guido and Drake. 2009. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace.

Väyrynen, Pekka. 2021. "Thick Ethical Concepts." In *The Stanford Encyclopedia of Philosophy,* Spring 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17:261–272. https://doi.org/10.1038/s41592-019-0686-2.

Walton, Douglas. 1997. *Appeal to expert opinion: Arguments from authority.* Pennsylvania State University Press.

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

Zollman, Kevin JS. 2007. "The communication structure of epistemic communities." *Philosophy of science* 74 (5): 574–587.

———. 2010. "The epistemic benefit of transient diversity." *Erkenntnis* 72 (1): 17–35.

7 Appendix

7.1 Contingency of Duijf's analytic results

With the restrictions $\kappa > \rho$, $\kappa > 0.5$ and $\rho > 0.5$ in place, Duijf derives further analytic results of his model, all but two of which break down if these restrictions are lifted:[33]

33. Duijf is, of course, aware of this dynamic, and this is clear from his paper.

$\frac{\partial p(D)}{\partial \kappa} = 1 - 2\rho$ shows that given perfect interest-alignment, the probability of disagreement ($p(D)$) between recipient and advisor decreases as the advisor's competency increases. This is contingent on the assumption that $\rho > 0.5$, as only this guarantees that $1 - 2\rho < 0$: The more competent an aligned advisor is, the higher the probability that they disagree with a recipient with a competency worse than chance. Similarly, $\frac{\partial p(D)}{\partial \rho} = 1 - 2\kappa$ is used to argue that given perfect interest alignment, the probability of disagreement also decreases as the recipient's competency increases. Given the assumption that $\kappa > \rho$, this decrease is even swifter than the above. However, this is again contingent on $1 - 2\kappa < 0$.

This is mirrored for the case of perfect anti-alignment: $\frac{\partial p(D)}{\partial \kappa} = 2\rho - 1$ shows that as advisor's competency increases, so does the chance of disagreement, and vice versa, $\frac{\partial p(D)}{\partial \rho} = 1 - 2\kappa$ shows that as the recipient's competency increases, so again does $p(D)$. Neither holds without the above restrictions: e.g., as their competency increases, the reliability of a consistently lying advisor goes down, and thus so does their probability of disagreement with a worse-than-chance recipient.

More generally, from $\frac{\partial p(D)}{\partial \rho} = (1-2\alpha)(2\kappa-1)$ Duijf concludes that as the probability of disagreement will decrease with increasing recipient competency iff the interests are aligned to more than 50%, and vice versa, from $\frac{\partial p(D)}{\partial \kappa} = (1 - 2\alpha)(2\rho - 1)$, he concludes same for increasing advisor competency. The former hinges on $\kappa > 0.5$, the latter on $\rho > 0.5$. He further concludes that increasing the degree of interest alignment always decreases the probability of disagreement from $\frac{\partial p(D)}{\partial \alpha} = 2(\kappa\overline{\rho} + \overline{\kappa}\rho) - 1$: given his assumptions, $(\kappa\overline{\rho} + \overline{\kappa}\rho)$ is guaranteed to be smaller than 0.5, making the right-hand side of this equation negative. Conceptually, however, it is clearly possible that increasing interest alignment leads to a higher chance of disagreement, for example between a knave advisor and a perfectly competent recipient.

And finally, concerning so-called regret, cases in which the recipient is correct whereas the advisor is not: from $\frac{\partial p(R)}{\partial \rho} = \alpha\overline{\kappa} + \overline{\alpha}\kappa$ Duijf concludes that as the recipient becomes more competent, the chance that they regret ($p(R)$) blindly deferring to their advisor increases. This does indeed hold even if we drop any restrictions on the values for κ, ρ, α. Additionally, the result that as an advisor's competency increases, the probability of regret goes down iff interests are aligned to more than 50% is maintained as well ($\frac{\partial p(R)}{\partial \kappa} = \rho(1 - 2\alpha)$). However, only given the assumption that $\kappa > 0.5$, it follows from $\frac{\partial p(R)}{\partial \alpha} = \rho(1 - 2\kappa)$ that as the degree of interest alignment increases, the probability of regret goes down. In cases of charlatan advisors that are less competent

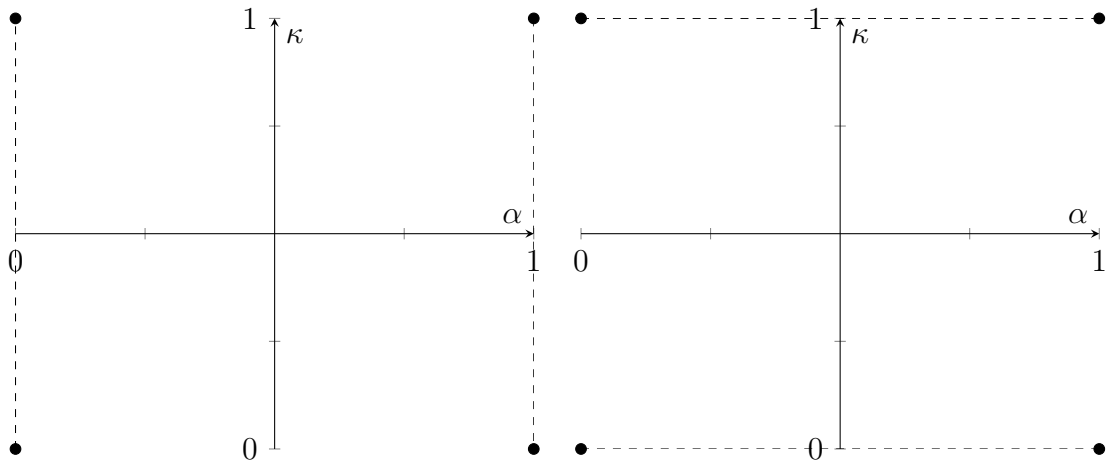

Figure 25: The half-way continuous SD-agent with competence being defined continuously and alignment as Boolean (left) versus with competency understood as Boolean and alignment as continuously defined (right).

than chance, increasing their alignment will instead decrease their overall reliability.

7.2 Half-way continuous focal models

For completeness sake, Figures 25 and 26 show how much of the testimonial situation SD-agents and AL-agents could conceptualize if we defined only one of their reliability-related source nodes as continuous, leaving the other Boolean. These models are otherwise not considered in this thesis, because once credence density functions are added to the model, it makes limited sense to do so for only one of two reliability-related nodes.

7.3 Updating the BH-agent

The following formulae allow calculation of which reports the Boolean BH-agent expects to receive from their advisor, as well as of how they will revise their beliefs about the hypothesis, and about the advisor's reliability, upon receiving positive or negative reports.

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

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

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

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

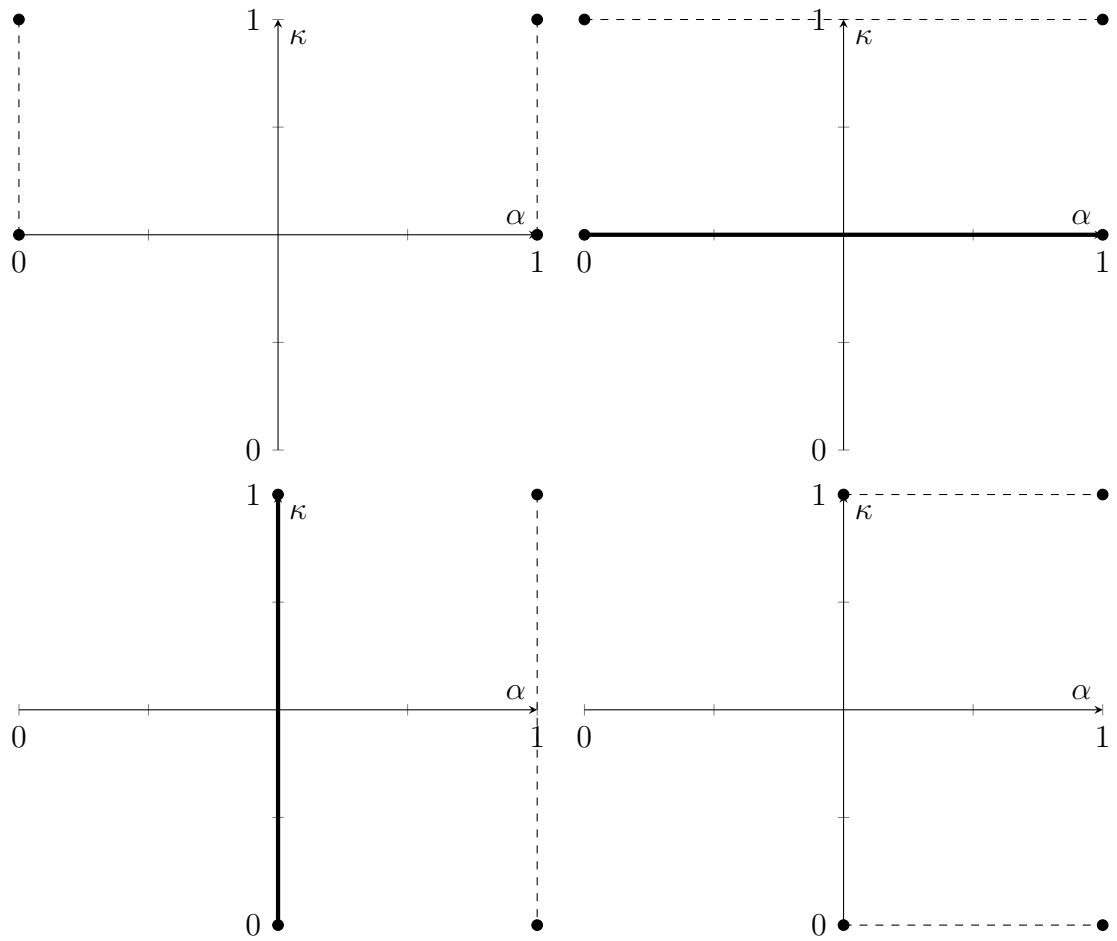

Figure 26: The half-way continuous AL-agent (above) and $AL\star$-agent (below) with competence being defined continuously and alignment as Boolean (left) versus with competency understood as Boolean and alignment as continuously defined (right).

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

To update $p(REP), p(HYP|REP)$ for the continuous BH-agent, replace the reliability credence value r by the expected value $M(r)$ of the reliability distribution. For trust revision, the continuous BH-agent needs to update their entire associated probability distribution $r(x)$ as follows.

(6) $r(x|REP) = \frac{p(REP|REL=x)r(x)}{p(REP)} = \frac{p(REP|REL)x+p(REP|\neg REL)\overline{x}}{p(REP)}r(x) = \frac{xh+\overline{x}\beta}{M(r)h+\beta-M(r)\beta}r(x)$ (from 1, Distribution updating)

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

7.4 Updating the AL-agent

The following formulae allow calculation of which reports the Boolean AL-agent expects to receive from their advisor, as well as of how they will revise their beliefs about the hypothesis, and their trust in the advisor's competency and alignment, upon receiving positive or negative reports.

(1) $p(REP) = hca + \overline{h}c\overline{a} + \beta\overline{c}$ (Law of total probability)

(2) $p(REP|HYP) = ca + \beta\overline{c}$ (from 1, for $h = 1$)

(3) $p(REP|COM) = ha + \overline{h}\overline{a}; p(REP|\neg COM) = \beta$ (from 1, setting $c = 1; c = 0$)

(4) $p(REP|AL) = hc + \beta\overline{c}; p(REP|\neg AL) = \overline{h}c + \beta\overline{c}$ (from 1, for $a = 1; a = 0$)

(5) $p(HYP|REP) = \frac{p(REP|HYP)p(HYP)}{p(REP)} = \frac{(ca+\beta\overline{c})h}{hca+\overline{h}c\overline{a}+\beta\overline{c}}$ (from 1, Bayes' rule)

(6) $p(HYP|\neg REP) = \frac{p(\neg REP|HYP)p(HYP)}{p(\neg REP)} = \frac{(1-(ca+\beta\overline{c}))h}{1-(hca+\overline{h}c\overline{a}+\beta\overline{c})}$ (from 5, Negation rule)

(7) $p(AL|REP) = \frac{p(REP|AL)p(AL)}{p(REP)} = \frac{(hc+\beta\overline{c})a}{hca+\overline{h}c\overline{a}+\beta\overline{c}}$ (from 1 and 3, Bayes' rule)

(8) $p(AL|\neg REP) = \frac{p(\neg REP|AL)p(AL)}{p(\neg REP)} = \frac{(1-(hc+\beta\overline{c}))a}{1-(hca+\overline{h}c\overline{a}+\beta\overline{c})}$(from 7, Negation rule)

(9) $p(COM|REP) = \frac{p(REP|COM)p(COM)}{p(REP)} = \frac{(ha+\overline{h}\overline{a})c}{hca+\overline{h}c\overline{a}+\beta\overline{c}}$ (from 1 and 4, Bayes' rule)

(10) $p(COM|\neg REP) = \frac{p(\neg REP|COM)p(COM)}{p(\neg REP)} = \frac{(1-(ha+\overline{h}\overline{a}))c}{1-(hca+\overline{h}c\overline{a}+\beta\overline{c})}$ (from 9, Negation rule)

To obtain the formulae for $p(REP)$ and $p(HYP|REP)$, for the continuous AL-agent, simply replace the credence-values for competency c and alignment a by the expected value of the respective distributions $M(c), M(a)$. However, for trust revision, the continuous AL-agent needs to simultaneously update their entire credence distributions for both source alignment and source competency.

(11) $a(x|REP) = \frac{p(REP|AL=x)a(x)}{p(REP)} = \frac{p(REP|AL)x+p(REP|\neg AL)\overline{x}}{p(REP)}a(x) = \frac{(hM(c)+\beta\overline{M(c)})x+(\overline{h}M(c)+\beta\overline{M(c)})\overline{x}}{hM(c)M(a)+\overline{h}M(c)\overline{M(a)}+\beta\overline{M(c)}}a(x)$ (from 1 and 4, Distribution updating)

(12) $a(x|\neg REP) = \frac{p(\neg REP|AL=x)a(x)}{p(\neg REP)} = \frac{p(\neg REP|AL)x+p(\neg REP|\neg AL)\overline{x}}{p(\neg REP)}a(x) = \frac{(1-(hM(c)+\beta\overline{M(c)}))x+(1-(\overline{h}M(c)+\beta\overline{M(c)}))\overline{x}}{1-(hM(c)M(a)+\overline{h}M(c)\overline{M(a)}+\beta\overline{M(c)})}a(x)$ (from 11, Negation rule)

(13) $c(x|REP) = \frac{p(REP|COM=x)c(x)}{p(REP)} = \frac{p(REP|COM)x+p(REP|\neg COM)\overline{x}}{p(REP)}c(x) = \frac{(hM(a)+\overline{hM(a)})x+\beta\overline{x}}{hM(c)M(a)+\overline{h}M(c)\overline{M(a)}+\beta\overline{M(c)}}c(x)$ (from 1 and 3, Distribution updating)

(14) $c(x|\neg REP) = \frac{p(\neg REP|COM=x)c(x)}{p(\neg REP)} = \frac{p(\neg REP|COM)x+p(\neg REP|\neg COM)\overline{x}}{p(\neg REP)}c(x) = \frac{(1-(hM(a)+\overline{hM(a)}))x+\overline{\beta}\overline{x}}{1-(hM(c)M(a)+\overline{h}M(c)\overline{M(a)}+\overline{\beta}\overline{M(c)})}$ (from 13, Negation rule)

7.5 Updating the SD-agent

The following formulae describe how the Boolean SD-agent revises their credence in the hypothesis, source alignment and source competency upon receiving positive or negative reports.

(1) $p(REP) = hca + \overline{hc}a + h\overline{ca} + \overline{h}c\overline{a}$ (Law of total probability)

(2) $p(REP|HYP) = ca + \overline{ca}$ (from 1, setting $h = 1$)

(3) $p(REP|AL) = hc + \overline{hc}; p(REP|\neg AL) = h\overline{c} + \overline{h}c$ (from 1, for $a = 1; a = 0$)

(4) $p(REP|COM) = ha + \overline{ha}; p(REP|\neg COM) = \overline{h}a + h\overline{a}$ (from 1, for $c = 1; c = 0$)

(5) $p(HYP|REP) = \frac{p(REP|HYP)p(HYP)}{p(REP)} = \frac{(ca+\overline{ca})h}{hca+\overline{h}\overline{c}a+h\overline{ca}+\overline{h}c\overline{a}}$ (from 1 and 2, Bayes' rule)

(6) $p(HYP|\neg REP) = \frac{p(\neg REP|HYP)p(HYP)}{p(\neg REP)} = \frac{(1-(ca+\overline{ca}))h}{1-(hca+\overline{h}\overline{c}a+h\overline{ca}+\overline{h}c\overline{a})}$ (from 5, Negation rule)

(7) $p(AL|REP) = \frac{p(REP|AL)p(AL)}{p(REP)} = \frac{(hc+\overline{hc})a}{hca+\overline{h}\overline{c}a+h\overline{ca}+\overline{h}c\overline{a}}$ (from 1 and 3, Bayes' rule)

(8) $p(AL|\neg REP) = \frac{p(\neg REP|AL)p(AL)}{p(\neg REP)} = \frac{(1-(hc+\overline{hc}))a}{1-(hca+\overline{h}\overline{c}a+h\overline{ca}+\overline{h}c\overline{a})}$(from 7, Negation rule)

(9) $p(COM|REP) = \frac{p(REP|COM)p(COM)}{p(REP)} = \frac{(ha+\overline{ha})c}{hca+\overline{h}ca+h\overline{c}a+\overline{h}c\overline{a}}$ (from 1 and 4, Bayes' rule)

(10) $p(COM|\neg REP) = \frac{p(\neg REP|COM)p(COM)}{p(\neg REP)} = \frac{(1-(ha+\overline{ha}))c}{1-(hca+\overline{h}ca+h\overline{c}a+\overline{h}c\overline{a})}$ (from 9, Negation rule)

As usual, to update $p(HYP)$ for the continuous version of the SD-agent instead, replace c, a by the means $M(c), M(a)$ of the respective probability distributions. To update these distributions themselves, use the following formulae:

(11) $a(x|REP) = \frac{p(REP|AL=x)a(x)}{p(REP)} = \frac{p(REP|AL)x+p(REP|\neg AL)\overline{x}}{p(REP)}a(x) = \frac{(hM(c)+\overline{hM(c)})x+(h\overline{M(c)}+\overline{h}M(c))\overline{x}}{hM(c)M(a)+\overline{h}\overline{M(c)}M(a)+h\overline{M(c)}\overline{M(a)}+\overline{h}M(c)\overline{M(a)}}a(x)$ (from 1 and 3, Distribution updating)

(12) $a(x|\neg REP) = \frac{p(\neg REP|AL=x)a(x)}{p(\neg REP)} = \frac{p(\neg REP|AL)x+p(\neg REP|\neg AL)\overline{x}}{p(\neg REP)}a(x) = \frac{(1-(hM(c)+\overline{hM(c)}))x+(1-(h\overline{M(c)}+\overline{h}M(c)))\overline{x}}{1-(hM(c)M(a)+\overline{h}\overline{M(c)}M(a)+h\overline{M(c)}\overline{M(a)}+\overline{h}M(c)\overline{M(a)})}a(x)$ (from 11, Negation rule)

(13) $c(x|REP) = \frac{p(REP|COM=x)c(x)}{p(REP)} = \frac{p(REP|COM)x+p(REP|\neg COM)\overline{x}}{p(REP)}c(x) = \frac{(hM(a)+\overline{hM(a)})x+(\overline{h}M(a)+h\overline{M(a)})\overline{x}}{hM(c)M(a)+\overline{h}\overline{M(c)}M(a)+h\overline{M(c)}\overline{M(a)}+\overline{h}M(c)\overline{M(a)}}c(x)$ (from 1 and 4, Distribution updating)

(14) $c(x|\neg REP) = \frac{p(\neg REP|COM=x)c(x)}{p(\neg REP)} = \frac{p(\neg REP|COM)x+p(\neg REP|\neg COM)\overline{x}}{p(\neg REP)}c(x) = \frac{(1-(hM(a)+\overline{hM(a)}))x+(1-(\overline{h}M(a)+h\overline{M(a)}))\overline{x}}{1-(hM(c)M(a)+\overline{h}\overline{M(c)}M(a)+h\overline{M(c)}\overline{M(a)}+\overline{h}M(c)\overline{M(a)})}c(x)$ (from 13, Negation rule)

7.6 Updating the OL-agent

The following formulae allow calculation of which reports the Boolean OL-agent expects to receive from their advisor, as well as of how they will revise their beliefs and trust upon receiving positive or negative reports.

(1) $p(REP) = rh + \overline{r}\overline{h}$ (Law of total probability)

(2) $p(HYP|REP) = \frac{p(REP|HYP)p(HYP)}{p(REP)} = \frac{rh}{rh+\overline{r}\overline{h}}$ (from 1, Bayes rule)

(3) $p(HYP|\neg REP) = \frac{p(\neg REP|HYP)p(HYP)}{p(\neg REP)} = \frac{\overline{r}h}{1-(rh+\overline{r}\overline{h})}$ (from 2, Negation rule)

(4) $p(REL|REP) = \frac{p(REP|REL)p(REL)}{p(REP)} = \frac{hr}{rh+\overline{r}\overline{h}}$ (from 1, Bayes rule)

(5) $p(REL|\neg REP) = \frac{p(\neg REP|REL)p(REL)}{p(\neg REP)} = \frac{\overline{h}r}{1-(rh+\overline{r}\overline{h})}$ (from 4, Negation rule)

To obtain the formulae for $c(REP)$ and $c(HYP|REP)$, for the continuous OL-agent, replace the credence in the reliability r by the expected value of the reliability distribution $M(r)$. For their trust revision ($p(REL|REP)$, $p(REL|\neg REP)$), the continuous OL-agent updates the the entire associated probability distribution $r(x)$ (see also Merdes, Sydow, and Hahn 2021, p. 5798).

(6) $r(x|REP) = \frac{p(REP|REL=x)r(x)}{p(REP)} = \frac{p(REP|REL)x+p(REP|\neg REL)\overline{x}}{p(REP)}r(x) = \frac{xh+\overline{x}\overline{h}}{M(r)h+\overline{M(r)\overline{h}}}r(x)$ (from 1, Distribution updating)

(7) $r(x|\neg REP) = \frac{p(\neg REP|REL=x)r(x)}{p(\neg REP)} = \frac{p(\neg REP|REL)x+p(\neg REP|\neg REL)\overline{x}}{p(\neg REP)}r(x) = \frac{x\overline{h}+\overline{x}h}{1-(M(r)h+\overline{M(r)\overline{h}})}r(x)$ (from 7, Negation rule)

7.7 Global updating for continuous models

I mention in Section 4.4 that continuous agents are capable of globally updating on inconsistent sets of reports from individual advisors. An in-depth exploration of these procedures and their results are beyond the scope of this thesis, but I will quickly discuss the simplest possible example, namely continuous OL-agents updating globally on two reports, using the DAG showcased in Figure 13. For this section, let E_1 be short for $\{REP_1, \neg REP_2\}$ and E_2 short for $\{REP_1, REP_2\}$. Due to its symmetric nature, the continuous OL-agent updates on all other combinations of two reports in mirrored or identical manners.

When updating locally with the continuous OL-agent, we were able to mostly re-use the formulae for $p(REP), p(REP|HYP), p(REP|REL)$, et cetera, by simply replacing every occurrence of $p(REL)$ with the weighted mean (expected value) of the reliability distribution $M(r)$. However, this is no longer possible for updating globally on more than a single report, because this would effectively remove the possibility for intermediary reliability values: for all four of the Boolean combinations of REL, HYP the conditional probability of E_1 is 0, making the Bayesian conditionalization on E_1 undefined for Boolean OL-agents.

The continuous OL-agent instead calculates $p(E)$ via the probabilistically weighted mean of $p(E|HYP, REL = x) + p(E|\neg HYP, REL = x)$ for all possible reliability values x. For E_1, each summand has the form $r(x)(hx\overline{x}+\overline{h}\overline{x}x)$, and for the 'consistent' E_2 it is $r(x)(hx^2 + \overline{h}\overline{x}^2)$ instead.

Globally updating on E_1 is achieved via $r(x|E_1) = \frac{p(E_1|REL=x)r(x)}{p(E_1)} = \frac{hx\overline{x}+\overline{h}\overline{x}x}{p(E_1)}r(x)$. To update on E_2, the continuous OL-agent proceeds as follows:
$r(x|E_2) = \frac{p(E_2|REL=x)r(x)}{p(E_2)} = \frac{hx^2+\overline{h}\overline{x}^2}{p(E_2)}r(x)$.

Just as one would expect, when a continuous OL-agent updates on E_1, their reliability distribution sharpens and its mean shifts towards 0.5, whereas it shifts towards 1 as a reaction to E_2 instead.

7.8 Computational implementation

I implemented these focal Bayesian agents in NetLogo 6.1.1 (Wilensky 1999) and, using its Python extension, out-sourced all computationally intensive calculations related to credence density distributions to Python 3.10.6 (Van Rossum 2009) and its SciPy (Virtanen et al. 2020) and NumPy (C. R. Harris et al. 2020) packages. This section contains a brief overview of the user interface (see Figure 27), as well as relevant code snippets[34]. I selected the OL-agent as the running example here, due to its relative simplicity. Keep in mind that like any other implementation of Bayesian reasoning, some compromises between accuracy and computational feasibility had to be made: floating point numbers in NetLogo are stored using 64 bits, and any resolution choice for storing the values of the credence density function is necessarily discrete.

The credence density functions for the continuous nodes are initialised using the beta function as follows, based on the resolution res of the unit interval. Here, r is a list storing the prior values of the reliability credence distribution. The choice of α or β values smaller than 1 leads to infinite credence density at $x = 0, x = 1$ respectively, and can therefore not be computed by this implementation.

```
(py:run
   "unitInterval = np.linspace(0, 1, (res + 1))"
   "r = stats.beta.pdf(unitInterval, alphaOL, betaOL)"
)
set r py:runresult "r"
```

This is how the expected value of a reliability distribution r with resolution res is calculated via the probabilistically weighted average, ...

```
(py:run
   "meanR = 0"
   "for x in range(len(r)):"
   "    meanR = meanR + r[x] * (x / res)"
   "meanR = meanR / (res + 1)"
)
set meanR py:runresult "meanR"
```

34. You can find the remainder of the model's code under `https://github.com/leon-schoeppl/source-reliability`.

... and via the slightly more accurate version of probabilistically weighing every value in the credence density distribution and then integrating it via trapezoid interpolation:[35]

```
(py:run
   "meanR = 0"
   "for x in range(len(r)):"
   "    r[x] = r[x] * (x / res)"
   "for x in range(len(r)-1):"
   "    meanR = meanR + (min(r[x], r[x+1]) * (1 / res) + 0.5 * abs(r[x] - r[x+1]) * (1 / res))"
)
set meanR py:runresult "meanR"
```

Here are the updating procedures of the Boolean OL-agent upon receiving a positive report (directly in NetLogo), followed by those for the continuous BH-agent (in Python, via NetLogo). h stores the credence in the hypothesis, and r either the reliability-credence, or its density distribution.

```
let tempH h
set h ((r * h)/(r * h + (1 - r)*(1 - h)))
set r ((tempH * r)/(tempH * r + (1 - tempH)*(1 - r)))
```

```
(py:run
   "for x in range(len(r)):"
   "    r[x] = ((x / res) * h + (1 - x / res) * (1 - h))/(meanR * h + (1 - meanR) * (1 - h)) * r[x]"
   "h = (meanR * h) / (meanR * h + (1 - meanR) * (1 - h))"
)
set r py:runresult "r"
set h py:runresult "h"
```

Finally, the veritistic value of each agent, and whether they accurately determined the advisor's trustworthiness (as described in Section 4.2), are updated every round as follows (continuous agents use $meanR$ in place of r):

```
set vValue ((phi * h) - (1 - phi) * (1 - h))

if (meanR > 0.5 and rel > 0.5) or (meanR < 0.5 and rel < 0.5) or (meanR = 0.5 and rel = 0.5) [
   set accurateTrustCounter accurateTrustCounter + 1
]
```

35. My choice of trapezoid interpolation and a resolution of 1000 is the result of an e-mail exchange with Christoph Merdes about the implementation of both Laputa, and of the models used in Merdes, Sydow, and Hahn 2021 and Hahn, Merdes, and Sydow 2018.

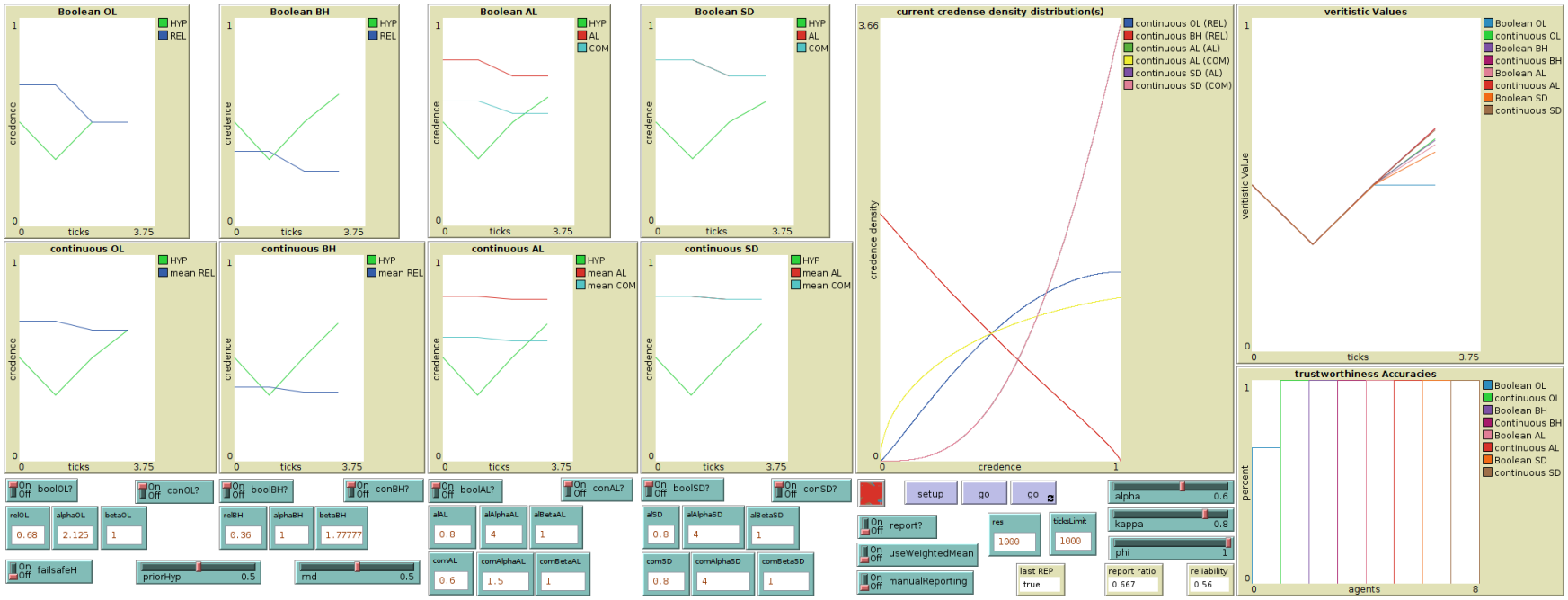

Figure 27: User interface for my implementation of the main eight Bayesian source-reliability models discussed in this thesis. The plots on the left monitor how the agent's (mean) credences develop as the simulation progresses. Below, are the inputs to toggle each individual agent on or off, as well as to determine their prior credences and credence distributions. The plots on the right monitor the exact shape of the credence density functions entertained by the continuously defined agents, how their veritistic values (see Section 2.3) develop over time, as well as their current trustworthiness accuracy scores (as defined in Section 4.2). Below are the inputs to determine, among other things, the resolution for the credence density functions and the objective values of the advisor's competency and degree of interest alignment. The model allows for reports to either be manually selected or stochastically generated based on the advisor's properties.

This specific screenshot shows the agents' behaviours upon receiving three reports (negative, positive, positive) from an advisor with $\alpha = 0.6, \kappa = 0.8$ about an objectively true φ. The focal agents' reliability priors are initialized as trusting, in accordance with Table 1, and their priors in the hypothesis are exactly 0.5.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig und ohne fremde Hilfe angefertigt habe. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autor*innen beruhen, sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, 20.02.2023 Name (+ Unterschrift)

This page was intentionally left blank.