Awareness Growth in Bayesian Networks

Reply to Steele and Stefánsson

Marcello/Rafal

1 Introduction

Learning is modeled in the Bayesian framework by the rule of conditionalization. This rule posits that the agent's new degree of belief in a proposition H after a learning experience E should be the same as the agent's old degree of belief in H conditional on E. That is,

$$\mathsf{P}^E(H) = \mathsf{P}(H|E),$$

where $\mathsf{P}()$ represents the agent's old degree of belief (before the learning experience E) and $\mathsf{P}^E()$ represents the agent's new degree of belief (after the learning experience E).

One assumption here is that E is learned with certainty. After the agent learns about E, there is no longer any doubt about the truth of E. This assumption has been the topic of extensive discussion in the literature.[1] The other assumption—which we focus on here—is that E and H belong to the agent's algebra of propositions. This algebra models the agent's awareness state, the propositions that the agent entertains as live possibilities.

Bayesian conditionalization makes it impossible for an agent to learn something they have never thought about. The algebra is fixed once and for all, since the learning experience does not modify it. This forces a great deal of rigidity on the learning process. Crucially, even before learning about E, the agent already knows the degree of belief in any proposition conditional on E. This picture commits the agent to the specification of their 'total possible future experience' (Howson 1976, The Development of Logical Probability), as though learning was confined to an 'initial prison' (Lakatos, 1968, Changes in the Problem of Inductive Logic).

But, arguably, the learning process is more complex than what conditionalization allows. Not only do we learn that some propositions that we were entertaining are true or false, but we may also learn new propositions that we did not entertain before. Or we may entertain new propositions—without necessarily learning that they are true or false—and this change in awareness may in turn change what we already believe. How should this more complex learning process be modeled by Bayesianism? Call this the problem of awareness growth. This problem can perhaps be divided into two parts: (i) how to model *learning* a new proposition not in the initial awareness state of the agent; (ii) how to model *entertaining* a new proposition not in the initial awareness state of the agent (without yet learning it).

Critics of Bayesianism and sympathizers alike have been discussing the problem of awareness growth under different names for quite some time, at least since the eighties. This problem arises in a number of different contexts, for example, new scientific theories (Glymour, 1980, Why I am not a Bayesian; Chihara 1987, Some Problems for Bayesian Confirmation Theory; Earmann 1992, Bayes of Bust?), language changes and paradigm shifts (Williamson 2003, Bayesianism and Language Change), and theories of induction (Zabell, Predicting the Unpredictable).

[1] As is well-known, Jeffrey's conditionalization relaxes this assumption.

Now, of course, the algebra of propositions could in principle be so rich to contain anything that could possibly be conceived, expressed, thought of. Such an algebra would not need to change at any point in the future. God-like agents could be associated with such rich algebra of propositions, but this is hardly a plausible model of ordinary agents with bounded resources such as ourselves. A fully comprehensive algebra of propositions cannot be the answer here.

A more promising proposal that has attracted considerable scholarly attention is Reverse Bayesianism (Karni and Viero, 2015, Probabilistic Sophistication and Reverse Bayesianism; Wenmackers and Romeijn 2016, New Theory About Old Evidence; Bradely 2017, Decision Theory with A Human Face) . The idea is to model awareness growth as a change in the algebra while ensuring that the probabilities of the propositions shared between the old and new algebra remain fixed under suitable constraints.

Let $\mathscr{F}$ be the initial algebra of propositions and let $\mathscr{F}^+$ the algebra after the agent's awareness has grown. Both contain the contradictory proposition $\bot$ and tautologous proposition $\top$ and they are closed under connectives such as disjunction $\vee$, conjunction $\wedge$ and negation $\neg$. Denote by X and X^+ the subsets of these algebras which contain only basic propositions, those that do not contain connectives. **Reverse Bayesianism** posits that the ratio of probabilities for any basic propositions A and B in both X and X^+—the basic propositions shared by the old and new algebra—remain constant through the process of awareness growth:

$$\frac{\mathsf{P}(A)}{\mathsf{P}(B)} = \frac{\mathsf{P}^+(A)}{\mathsf{P}^+(B)},$$

where $\mathsf{P}()$ represents the agent's degree of belief before awareness growth and $\mathsf{P}^+()$ represents the agent's degree of belief after awareness growth.

What is the justification for Reverse Bayesianism? Perhaps the best justification is pragmatic. As the awareness of an an agent grows, the agent would prefer not to throw away completely the epistemic work they have done so far. The agent may desire to retain as much of their old degrees of beliefs as possible. Reverse Bayesianism provides a simple recipe to do that. It also coheres with the conservative spirit of conditionalization which preserves the old probability distribution conditional on what is learned.

Reverse Bayesianism is an elegant theory that manages to cope with a seemingly intractable problem. But, unfortunately, it is not without complications. Steele and Stefánsson (2021, Belief Revision for Growing Awareness) argue that Reverse Bayesianism, when suitably formulated, can work in a limited class of cases, what they call *awareness expansion*, but cannot work for *awareness refinement* (more on this distinction later). Their argument rests on a number of ingenious counterexamples. We contend, however, that their counterexamples have limited applicability and thus constitute an overall weak case against Reverse Bayesianism. Our argument relies on the theory of Bayesian networks and makes two key claims. First, besides cases of expansion, there are also cases of refinement in which Reverse Bayesianism can be made to work. Second, Steele and Stefánsson's counterexamples only target a circumscribed class of refinement cases, and even for those, a conservative constraint in the spirit of Reverse Bayesianism can be carved out.

2 Counterexamples?

We begin by rehearsing two of the ingenious counterexamples to Reverse Bayesianism by Steele and Stefánsson. One targets awareness expansion and the other awareness refinement. The difference is intuitively plausible, but it can be tricky to pin down formally. A rough characterization will suffice here. Suppose, as is customary, propositions are interpreted as sets of possible worlds, where the set of all possible worlds is the possibility space. An algebra of

propositions thus interpreted induces a partition of the possibility space. Refinement occurs when the new proposition added to the algebra induces a more fine-grained partition of the possibility space. Expansion, instead, occurs when the new proposition shows the existing partition of the possibility space is not exhaustive and more possible worlds are added.

The first counterexample by Steele and Stefánsson targets cases of awareness expansion:

> FRIENDS: Suppose you happen to see your partner enter your best friend's house on an evening when your partner had told you she would have to work late. At that point, you become convinced that your partner and best friend are having an affair, as opposed to their being warm friends or mere acquaintances. You discuss your suspicion with another friend of yours, who points out that perhaps they were meeting to plan a surprise party to celebrate your upcoming birthday—a possibility that you had not even entertained. Becoming aware of this possible explanation for your partner's behaviour makes you doubt that she is having an affair with your friend, relative, for instance, to their being warm friends. (Steele and Stefánsson, 2021, Section 5, Example 2)

Why does the scenario FRIENDS conflict with Reverse Bayesianim? Even though Steele and Stefánsson do not provide the details, it pays to be explicit here at the cost of pedantry.

Initially, the algebra only contains the hypotheses 'my partner and my best friend met to have an affair' (*Affair*) and 'my partner and my best friend met as friends or acquaintances' (*Friends/acquaintances*). The other proposition in the algebra is the evidence at your disposal, that is, the fact that your partner and your best friend met one night secretively and without telling you (*Secretive*). There may be other propositions, but these are the ones to focus on. Hypothesis *Affair* better explains the evidence at your disposal than hypothesis *Friends/acquaintances*. In probabilistic terms, this can be expressed by comparing likelihoods:

$$\mathsf{P}(\textit{Secretive}|\textit{Affair}) > \mathsf{P}(\textit{Secretive}|\textit{Friends/acquaintances}),$$

from which it also follows that *Affair* is more probable than *Friends/acquaintances*

$$\mathsf{P}(\textit{Affair}|\textit{Secretive}) > \mathsf{P}(\textit{Friends/acquaintances}|\textit{Secretive}), \tag{>}$$

so long as the prior probabilities of the two hypotheses are not skewed in one direction.[2]

Next, the algebra changes. A new hypothesis is added which you had not considered before: your partner and your best friends met to plan a surprise party for your upcoming birthday (*Surprise*). This is a game changer. The evidence *Secretive* now makes better sense in light of this new hypothesis than the hypothesis *Affair*:

$$\mathsf{P}^{+}(\textit{Secretive}|\textit{Surprise}) > \mathsf{P}^{+}(\textit{Secretive}|\textit{Affair}).$$

And, this new hypothesis should be more likely than the hypothesis *Affair*:

$$\mathsf{P}^{+}(\textit{Surprise}|\textit{Secretive}) > \mathsf{P}^{+}(\textit{Affair}|\textit{Secretive}). \tag{*}$$

Reverse Bayesianism is not yet in trouble. Steele and Stefánsson, however, conclude that the probability of *Friends/acquaintances* should now exceed that of *Affair*. They write: 'Becoming aware of this possible explanation for your partner's behaviour makes you doubt that she is having an affair with your friend, relative, for instance, to their being warm friends.' So:

$$\mathsf{P}^{+}(\textit{Affair}|\textit{Secretive}) < \mathsf{P}^{+}(\textit{Friends/acquaintances}|\textit{Secretive}). \tag{<}$$

[2] If you were initially nearly certain your partner could not possibly have an affair, even the fact they behaved very secretively or lied to you might not affect the probability of the two hypotheses.

Arguably, this holds because *Surprise* implies *Friends/acquaintances*. In order to prepare a surprise party, your partner and best friend have to be at least acquaintances. And given that one implies the other, if *Surprise* is more likely than *Affair* (by $*$), then *Friends/acquaintances* must also be more likely than *Affair*. And if both ($>$) and ($<$) holds, the ratio of the probabilities of basic propositions is not fixed before and after the episode of awareness growth. This is a violation of Reverse Bayesianism.

But, as Steele and Stefánsson admits, Reverse Bayesianism is not really in trouble here. It can still be made to work with a slightly different—though quite similar in spirit—condition, called **Awareness Rigidity**:

$$\mathsf{P}^{+}(A|T^{*}) = \mathsf{P}(A),$$

where T^{*} corresponds to a proposition that picks out, from the vantage point of the new awareness state, what corresponds to the entire possibility space before the episode of awareness growth. In our running example, the proposition $\neg$*Surprise* picks out the entire possibility space before the episode of awareness growth. So Awareness Rigidity would require that:

$$\mathsf{P}^{+}(\textit{Friends/acquaintances}|\neg\textit{Surprise}) = \mathsf{P}(\textit{Friends/acquaintances}).$$

Conditional on $\neg$*Surprise*, it is indeed true that the probability of *Friends/acquaintances* has not changed before and after the episode of awareness growth. And it is also true that *Affair* remains the most likely hypothesis in light of the evidence (again conditional on $\neg$*Surprise*):

$$\mathsf{P}^{+}(\textit{Affair}|\textit{Secretive}\&\neg\textit{Surprise}) > \mathsf{P}^{+}(\textit{Friends/acquaintances}|\textit{Secretive}\&\neg\textit{Surprise}).$$

Awareness Rigidity is vindicated. Reverse Bayesianism—the spirit of it, not the letter—stands.

This is not the end of the story, however. Steele and Stefánsson offer another counterexample to Reverse Bayesianism (which also works against Awareness Rigidity):

> MOVIES: Suppose you are deciding whether to see a movie at your local cinema. You know that the movie's predominant language and genre will affect your viewing experience. The possible languages you consider are French and German and the genres you consider are thriller and comedy. But then you realise that, due to your poor French and German skills, your enjoyment of the movie will also depend on the level of difficulty of the language. Since it occurs to you that the owner of the cinema is quite simple-minded, you are, after this realisation, much more confident that the movie will have low-level language than high-level language. Moreover, since you associate low-level language with thrillers, this makes you more confident than you were before that the movie on offer is a thriller as opposed to a comedy. (Steele and Stefánsson, 2021, Section 5, Example 3)

Initially, you did not consider language difficulty. The algebra contained the propositions *French* and *German*, as well as *Thriller* and *Comedy*. Then, you realized another variable might be at play, namely the level of difficulty of the language of the movie. The realization that the owner is simple-minded suggests that the level of linguistic difficulty of the movie will be low. The latter in turn suggests that the movie is more likely a thriller rather than a comedy (perhaps because thrillers are simpler—linguistically—than comedies). So, against Reverse Bayesianism, the scenario MOVIES violates the condition $\frac{\mathsf{P}(\textit{Thriller})}{\mathsf{P}(\textit{Comedy})} = \frac{\mathsf{P}^{+}(\textit{Thriller})}{\mathsf{P}^{+}(\textit{Comedy})}$.

The counterexample also works against Awareness Rigidity. It is not true that $\mathsf{P}(\textit{Thriller}) = \mathsf{P}^{+}(\textit{Thriller}|\textit{Thriller} \vee \textit{Comedy})$. To see why that is, note that this counterexample is a case of refinement. First, you categorize movies by just language and genre, and then you add a further category, level of difficulty. In the scenario FRIENDS, instead, the possibility space grew by adding situations in which your partner and best friends met neither as lovers nor

solely as friends.[3] So if MOVIES is a case of refinement, the proposition which picks out the entire possibility space should be the same before and after awareness growth, for example, *Thriller* $\vee$ *Comedy*. In cases of awareness growth by refinement, then, Awareness Rigidity mandates that all probability assignments stay the same.

How strong of a counterexample is MOVIES? For the sake of clarity, it can be split into two episodes. In the first, you entertain a new variable besides language and genre, namely the language difficulty of the movie. In the second episode, you learn something you did not consider before, namely that the owner is simple-minded. The first is a case of mere refinement: you simply entertain a new way of categorizing movies. The second is a case of learning: you learn something you did not consider before. The tacit assumption is that awareness growth is both *entertaining* a new proposition not in the initial awareness state of the agent (without learning it) and *learning* a new proposition not in the initial awareness state of the agent. We agree with Steele and Stefánsson that both these phenomena should count as instances of awareness growth. But one wonders. Is the second episode (learning something new) necessary for the counterexample to work together with the first episode (mere refinement)?

Suppose the counterexample only works in tandem with an episode of learning something new. In that were so, defenders of Reverse Bayesianism or Awareness Rigidity could still claim that their theory applies to a large class of cases. It applies to cases of awareness refinement without learning and also to cases of awareness expansion. For recall that the first putative counterexample about awareness expansion—FRIENDS—did not challenge Reverse Bayesianism insofar as the latter is formulated in terms of its close cousin, Awareness Rigidity. So the force of Steele and Stefánsson's counterexamples would be rather limited.

Or perhaps there is a more straightforward counterexample that only depicts mere refinement without an episode of learning and that still challenges Reverse Bayesianism and Awareness Rigidity? As we shall soon see, the answer to this question is indeed positive.

3 A simpler counterexample

Steele and Stefánsson's counterexample to Reverse Bayesianism in the case of refinement is rather complex, perhaps unnecessarily so. We now present something simpler:

> LIGHTING: You have evidence that favors a certain hypothesis, say a witness saw the defendant around the crime scene. You give some weight to this evidence. In your assessment, that the defendant was seen around the crime scene raises the probability that the defendant was actually there. But now you wonder, what if it was dark when the witness saw the defendant? You become a bit more careful and settle on this: if the lighting conditions were good, you should still trust the evidence, but if they were bad, you should not. Unfortunately, you cannot learn about the actual lighting conditions, but the mere realization that it *could* have been dark makes you change the probability that the defendant was actually there, based on the same eveidence.

This scenario is simpler because it consists of mere refinement. You wonder about the lighting conditions but you do not learn what they were. Still, mere refinement in this scenario challenges Reverse Bayesianism and Awareness Rigidity. That this should be so is not easy to see. Fortunately, the theory of Bayesian networks helps to see why.

A Bayesian network is a formal model that consists of a graph accompanied by a probability

[3] The addition of the hypothesis *Surprise* is, however, an ambiguous case. For one thing, *Surprise* is a novel hypothesis that cannot be subsumed under *Friends/acquaintances* or *Affair*. On the other, *Surprise* seems a refinement of *Friends/acquaintances*, since a meeting for planning a surprise is a more specific way to describe a meeting of acquaintances. We will provide a more clear-cur example of expansion later in the paper.

distribution. The nodes in the graph represent random variables that can take different values. We will use 'nodes' and 'variables' interchangeably. The nodes are connected by arrows, but no loops are allowed, hence the name direct acyclic graph (DAG). In this framework, awareness growth brings about a change in the graphical network—nodes and arrows are added or erased—as well as a change in the probability distribution from the old to the new network.

To model the scenario LIGHTING with Bayesian networks, we start with this graph:

$$H \rightarrow E,$$

where H is the hypothesis node and E the evidence node. If an arrow goes from H to E, the probability distribution associated with the Bayesian network should be defined by conditional probabilities of the form $\mathsf{P}(E = e|H = h)$, where uppercase letters represent the variables (nodes) and lower case letters represent the values of these variables.[4]

Since you trust the evidence, you think that it is more likely under the hypothesis that the defendant was present at the crime scene than under the alternative hypothesis:

$$\mathsf{P}(E{=}seen|H{=}present) > \mathsf{P}(E{=}seen|H{=}absent)$$

It is not necessary to fix exact numerical values for these conditional probabilities. The inequality is a qualitative ordering of how plausible the evidence is in light of competing hypotheses. No matter the numbers, by the probability calculus, it follows that the evidence raises the probability of the hypothesis $H{=}present$:

$$\mathsf{P}(H{=}present|E{=}seen) > \mathsf{P}(H{=}present)$$

Now, as you wonder about the lighting conditions, the graph should be amended:

$$H \rightarrow E \leftarrow L,$$

where the node L can have two values, $L{=}good$ and $L{=}bad$. A plausible way to update your assessment of the evidence is as follows:

$$\mathsf{P}^{+}(E{=}seen|H{=}present \wedge L{=}good) > \mathsf{P}^{+}(E{=}seen|H{=}absent \wedge L{=}good)$$

$$\mathsf{P}^{+}(E{=}seen|H{=}present \wedge bad) = \mathsf{P}^{+}(E{=}seen|H{=}absent \wedge L{=}bad)$$

Note the change in the probability function from $\mathsf{P}()$ to $\mathsf{P}^{+}()$. Here is what you are thinking: if the lighting conditions were good, you should still trust the evidence like you did before. But if the lighting conditions were bad, you should regard the evidence as no better than chance. Again, there are no exact numerical values here.

Should you now assess the evidence at your disposal—that the witness saw the defendant at the crime scene—any differently than you did before? Would it be wrong to think the evidence had the same value? The evidence would have the same value if the likelihood ratios associated with it relative to the competing hypotheses were the same before and after awareness growth:

$$\frac{\mathsf{P}(E = e|H = h)}{\mathsf{P}(E = e|H = h')} = \frac{\mathsf{P}^{+}(E = e|H = h)}{\mathsf{P}^{+}(E = e|H = h')}. \tag{C}$$

But it would be quite a coincidence if (C) were true. For concreteness, let's use some numbers:

$$\mathsf{P}(E{=}seen|H{=}present) = \mathsf{P}^{+}(E{=}seen|H{=}present \wedge L{=}good) = .8$$

[4] A major point of contention in the interpretation of Bayesian networks is is the meaning of the directed arrows. They could be interpreted causally—as though the direction of causality proceeds from the events described by the hypothesis to event described by the evidence—but they need not be. REFERENCES?

$$\mathsf{P}(E{=}seen|H{=}absent) = \mathsf{P}^{+}(E{=}seen|H{=}absent \wedge L{=}good) = .4$$
$$\mathsf{P}^{+}(E{=}seen|H{=}present \wedge L{=}bad) = \mathsf{P}^{+}(E{=}seen|H{=}absent \wedge L{=}bad) = .5.$$

So the ratio $\frac{\mathsf{P}(E=seen|H=present)}{\mathsf{P}(E=seen|H=absent)} = 2$. Before awareness growth, you thought the evidence favored the hypothesis *H=present* moderately strongly. That seemed reasonable. But, after the awareness growth, the ratio $\frac{\mathsf{P}^{+}(E=seen|H=present)}{\mathsf{P}^{+}(E=seen|H=absent)} = \frac{.65}{.45} \approx 1.44$.[5] This argument can be repeated with several other numerical assignments. So, quite often, mere refinement can weaken the evidence, even without learning anything new. Of course, if you did learn that the lighting conditions were bad, the evidence would become even weaker, effectively worthless:

Need a more general argument here. Simulation?

$$\frac{\mathsf{P}^{+,L=bad}(E{=}seen|H{=}present)}{\mathsf{P}^{+,L=bad}(E{=}seen|H{=}absent)} = 1,$$

where $\mathsf{P}^{+,L=bad}()$ is the new probability function after learning that *L=bad*.

Why does all this matter? We have seen that, after awareness growth, you should regard the evidence at your disposal as one that favors *H=present* less strongly or not at all. Since the prior probability of the hypothesis should be the same before and after awareness growth, it follows that

$$\mathsf{P}^{+}(H{=}present|E{=}seen) \neq \mathsf{P}(H{=}present|E{=}seen).$$

This outcome violates Awareness Rigidity. For recall that in cases of refinement, Awareness Rigidity requires that the probability of basic propositions stay fixed.

Reverse Bayesianism is also violated. For example, the ratio of the probabilities of *H=present* to *E=seen*, before and after awareness growth, has changed:

$$\frac{\mathsf{P}^{E=seen}(H{=}present)}{\mathsf{P}^{E=seen}(H{=}seen)} \neq \frac{\mathsf{P}^{+,E=seen}(H{=}present)}{\mathsf{P}^{+,E=seen}(H{=}seen)},$$

where $\mathsf{P}^{E=seen}()$ and $\mathsf{P}^{+,E=seen}()$ represent the agent's degrees of belief, before and after awareness growth, updated by the evidence *E=seen*.

All in all, the counterexample LIGHTING works even though it only depicts refinement without learning, and thus strengthens Steele and Stefánsson's case against Reverse Bayesianism and Awareness Rigidity. But despite that, their criticism still remains more circumscribed that it might appear at first. There are cases of refinement in which Reverse Bayesianism and Awareness Rigidity are perfectly fine in their place. This is our next topic.

[5] The calculations here rely on the dependency structure encoded in the Bayesian network (see starred step below).

$$\begin{aligned}
\mathsf{P}^{+}(E{=}seen|H{=}present) &= \mathsf{P}^{+}(E{=}seen \wedge L{=}good|H{=}present) + \mathsf{P}^{+}(E{=}seen \wedge L{=}bad|H{=}present) \\
&= \mathsf{P}^{+}(E{=}seen|H{=}present \wedge L{=}good) \times \mathsf{P}^{+}(L{=}good|H{=}present) \\
&+ \mathsf{P}^{+}(E{=}seen|H{=}present \wedge L{=}bad) \times \mathsf{P}^{+}(L{=}bad|H{=}present) \\
&=^{*} \mathsf{P}^{+}(E{=}seen|H{=}present \wedge L{=}good) \times \mathsf{P}^{+}(L{=}good) \\
&+ \mathsf{P}^{+}(E{=}seen|H{=}present \wedge L{=}bad) \times \mathsf{P}^{+}(L{=}bad) \\
&= .8 \times .5 + .5 * .5 = .65
\end{aligned}$$

$$\begin{aligned}
\mathsf{P}^{+}(E{=}seen|H{=}absent) &= \mathsf{P}^{+}(E{=}seen \wedge L{=}good|H{=}absent) + \mathsf{P}^{+}(E{=}seen \wedge L{=}bad|H{=}absent) \\
&= \mathsf{P}^{+}(E{=}seen|H{=}absent \wedge L{=}good) \times \mathsf{P}^{+}(L{=}good|H{=}absent) \\
&+ \mathsf{P}^{+}(E{=}seen|H{=}absent \wedge L{=}bad) \times \mathsf{P}^{+}(L{=}bad|H{=}absent) \\
&=^{*} \mathsf{P}^{+}(E{=}seen|H{=}absent \wedge L{=}good) \times \mathsf{P}^{+}(L{=}good) \\
&+ \mathsf{P}^{+}(E{=}seen|H{=}absent \wedge L{=}bad) \times \mathsf{P}^{+}(L{=}bad) \\
&= .4 \times .5 + .5 * .5 = .45
\end{aligned}$$

4 Another refinement

Consider this variation of the LIGHTING scenario:

> VERACITY: A witness saw that the defendant was around the crime scene and you initially took this to be evidence that the witness was actually there. But then you had second thoughts. Instead of worrying about the lighting conditions, you worry that the witness might be lying or misremembering what happened. Perhaps, the witness was never there, made things up or mixed things up. But despite that, you do not change anything of your initial assessment of the evidence.

The rational thing to do here is to stick to your guns and not change your earlier assessment of the evidence. Why should that be so? And what is the difference with LIGHTING? Once again, Bayesian networks proves to be a good analytic tool here.

The graphical network should initially look like this:

$$H \rightarrow E$$

But, as your awareness grows, the graphical network should be updated:

$$H \rightarrow E \rightarrow R$$

The hypothesis node H bears on the whereabouts of the defendant. Note the difference between E and R. The evidence node bears on what the witness saw. The reporting node bears on what the witness reports to have seen. The chain of transmission from 'seeing' to 'reporting' may fail for various reasons, such as lying or confusion.

It pays to highlight the difference between LIGHTING and VERACITY. They are both cases of refinement. In one, what the witness saw could have occurred under good or bad lighting conditions; in the other, what the witness saw could have been reported truthfully or untruthfully. But refinement is structurally different in the two cases. In LIGHTING, the connection between the evidence and the hypothesis undergoes a change, since the lighting conditions affect the witness' ability to have reliable experiences of what happened. In VERACITY, instead, the connection between the evidence and the hypothesis is not affected. At stake is the extent to which what the witness saw, if anything, is reported truthfully or not.

So, even if VERACITY is a case of refinement, the old and new probability functions agree with one another completely. The conditional probabilities, $\mathsf{P}(E = e|H = h)$ should be the same as $\mathsf{P}^{+}(E = e|H = h)$ for any values e and h of the variables H and E that are shared before and after awareness growth. Given the dependency structure of the two Bayesian networks—first, $H \rightarrow E$ and then $H \rightarrow E \rightarrow R$—the equality is easy to establish formally.[6] Thus, Reverse Bayesianism and Awareness Rigidity are perfectly fine in scenarios like VERACITY.

A confusion should be eliminated at this point. We do not intend to suggest that the assessment of the probability of the hypothesis *H=present* should undergo no change at all. If you worry that the witness could have lied, shouldn't this affect your degree of beliefs in the veracity of what they said about the defendant's whereabouts? Surely so. To see where the confusion might lie, note that in VERACITY an episode of awareness refinement may actually take place together with a form of retraction. Initially, what is taken to be known, after the learning episode, is that the witness *saw* the defendant around the crime scene. But after awareness growth, you realize your learning is in fact limited to what the witness *reported* to have seen. So the previous learning episode is retracted and replaced by a more careful statement of what you learned. This retraction will affect the probability you assign to the hypothesis *H=seen*, but this does not conflict with Reverse Bayesianism or Awareness Rigidity.

[6] GIVE PROOF

In LIGHTING, instead, no retraction of the evidence takes place. The evidence that is known remains the fact that the witness saw the defendant around the crime scene, even though that experience could have been misleading due to bad lighting conditions.

Where does this leave us? The following are now well-established: (a) Reverse Bayesianism (or its close cousin Awareness Rigidity) handles successfully cases of awareness expansion; (b) it also handles successfully cases of refinement like VERACITY; but (c) it does fail in cases of refinement like LIGHTING. Bayesian networks helped to distinguish these two forms of refinement, and there may be other, more fine-grained distinctions to be made. So, ultimately, Steele and Stefánsson's argument only targets a subclass of refinement cases. But even in those—we shall soon see—there is room to carve out a conservative constraint that is close in spirit to Reverse Bayesianism.

5 A modest conservative constraint

Recall that, in LIGHTING, the probability functions $\mathsf{P}()$ and $\mathsf{P}^+()$ do not assign the same weight to the evidence relative to the competing hypotheses, except in somewhat exceptional circumstances. Hence, Reverse Bayesianism and Awareness Rigidity fail in this scenario. But despite that, the two probability functions agree in one important respect:

$$\mathsf{P}(E = e|H = h) \geq \mathsf{P}(E = e|H = h') \textit{ iff } \mathsf{P}^+(E = e|H = h) \geq \mathsf{P}^+(E = e|H = h'), \qquad (C^*)$$

where (i) E and H are nodes that are part of the graphical network before and after awareness growth, and (ii) there is a direct path from H to E before and after awareness growth. In other words, the plausibility ordering between hypotheses and evidence is preserved. Condition (C^*) can serve as a conservative constraint that governs the relationship between $\mathsf{P}()$ and $\mathsf{P}^+()$. It is satisfied in the scenario LIGHTING, but how general is this condition?

Is the condition of direct path necessary?

We now show that (C^*) holds generally in a class of Bayesian networks, under minimal, and entirely reasonable, assumptions. Assume the Bayesian network has a node E with an incoming arrow from node H, before and after awareness growth. After awareness growth, besides E and H, another variable Y is under consideration. The new graph looks like this:

$$H \rightarrow E \leftarrow Y.$$

For simplicity, we assume that variables are binaries. All we need is the following assumption:

$$\mathsf{P}(E = e|H = h) \geq \mathsf{P}(E = e|H = h')$$

$$\textit{iff}$$

$$\mathsf{P}^+(E = e|H = h \wedge Y = y) \geq \mathsf{P}^+(E = e|H = h' \wedge Y = y) \qquad \text{(EQUAL)}$$

$$\textit{iff}$$

$$\mathsf{P}^+(E = e|H = h \wedge Y = y') \geq \mathsf{P}^+(E = e|H = h' \wedge Y = y')$$

This assumption says that the plausibility ordering remains the same before and after awareness growth *all else being the same*. It is a minimal assumption, but enough to establish (C^*).[7]

[7] From (EQUAL) and via this chain of equivalences:

$$[a \geq a' \& b \geq b'] \textit{ iff } [ak \geq a'k \& b(1-k) \geq b'(1-k) (\textit{with } k > 0)] \textit{ iff } [ak + b(1-k) \geq a'k + b'(1-k)],$$

it follows that

$$\mathsf{P}(E = e|H = h) \geq \mathsf{P}(E = e|H = h')$$

$$\textit{iff}$$

There may be cases in which the plausibility ordering is not preserved because (EQUAL) does not hold. This would be cases of *transformative* awareness growth. For suppose you have evidence that—in your judgment—reliably tracks a hypothesis, say you think that appearance as of hands reliably tracks the presence of hands:

$$\mathsf{P}(E = \textit{as-of-hands} | H = \textit{hands}) > \mathsf{P}(E = \textit{as-of-hands} | H = \textit{no-hands})$$

You now entertain a 'switching hypothesis' in which a demon switches things around: when you see a hand, there is no hand, and when you do not see a hand, there is a hand. In this case, (EQUAL) would be violated since

$$\mathsf{P}^+(E = \textit{as-of-hands} | H = \textit{hands} \wedge Y = \textit{switching}) < \mathsf{P}^+(E = \textit{as-of-hands} | H = \textit{no-hands} \wedge Y = \textit{switching})$$

In more common skeptical scenarios without a switching hypothesis, however, (EQUAL) should still hold. For what skeptical scenarios often do is to neutralize the value of the evidence, not so much switching its value.

The preservation of the plausibility ordering formulated by condition (C^*) should also hold in Steele and Stefánsson's scenario MOVIES. We now briefly explain why. At first, the graphical network looks like this:

$$\textit{Genre} \rightarrow \textit{Enjoyment} \leftarrow \textit{Language},$$

where each node can take two values: *Genre=comedy* and *Genre=thriller*; *Language=french* and *Language=german*; and *Enjoyment=yes* and *Enjoyment=no*. Assume you are ranking the options in terms of how they are going to contribute to your enjoyment (*Enjoyment=yes*). You are more likely to enjoy a comedy in French than anything else, but you are more likely to enjoy a thriller in German than one in French, and your lowest preference is for a comedy in German. This ranking can be encoded by conditional probability statements of the form

$$\mathsf{P}(\textit{Enjoyment=}x | \textit{Language=}y \wedge \textit{Genre=}z) \geq \mathsf{P}(\textit{Enjoyment=}x | \textit{Language=}y' \wedge \textit{Genre=}z').$$

The first episode of awareness growth in MOVIES consists in realizing that the linguistic difficulty of the movie could also be a factor. So the expanded graphical network now becomes:

$$\begin{array}{c} \textit{Difficulty} \\ \downarrow \\ \textit{Genre} \rightarrow \textit{Enjoyment} \leftarrow \textit{Language} \end{array}$$

Your ranking of what is likely to give you enjoyment should now be updated and made more specific, but much of the earlier ordering can be retained, that is:

$$\mathsf{P}(\textit{Enjoyment=}x | \textit{Language=}y \wedge \textit{Genre=}z) \geq \mathsf{P}(\textit{Enjoyment=}x | \textit{Language=}y' \wedge \textit{Genre=}z')$$

iff

$$\mathsf{P}^+(\textit{Enjoyment=}x | \textit{Language=}y \wedge \textit{Genre=}z) \geq \mathsf{P}^+(\textit{Enjoyment=}x | \textit{Language=}y' \wedge \textit{Genre=}z').$$

$$\mathsf{P}^+(E = e | H = h \wedge Y = y) \times \mathsf{P}^+(Y = y) + \mathsf{P}^+(E = e | H = h \wedge Y = y') \times \mathsf{P}^+(Y = y')$$
$$\geq$$
$$\mathsf{P}^+(E = e | H = h' \wedge Y = y) \times \mathsf{P}^+(Y = y) + \mathsf{P}^+(E = e | H = h' \wedge Y = y') \times \mathsf{P}^+(Y = y')$$

We are done since, by the law of total probability and the probabilistic dependencies in the graph, (C^*) is equivalent to the above statement.

The difference with condition (C^*) is that here two propositions, not just one, are conditioned on. So (C^*) should be be generalized, accordingly, but the general idea remains the same.[8]

6 Conclusion

We argued that the case against Reverse Bayesianism is much weaker than one might think. The scenario MOVIES—which is Steele and Stefánsson's key counterexample to Reverse Bayesianism—is unconvincing since it mixes learning and refinement. To avoid this, we constructed a more clear-cut case of refinement, LIGHTING, in which both Awareness Rigidity and Reverse Bayesianism fail unequivocally. However, we showed that even in cases like this a modest conservative constraint can still be carved out, vindicating to some extent the spirit of Reverse Bayesianism. We also showed that there are cases of refinement like VERACITY in which Reverse Bayesianism and Awareness Rigidity are perfectly fine in their place.

We conclude with a few programmatic observations. We think that the awareness of agents grows while holding fixed certain material structural assumptions, based on commonsense, semantic stipulations or causal dependency. To model awareness growth, we need a formalism that can express these material structural assumptions. This can done using Bayesian networks, and we offered some illustrations of this strategy, for example, by distinguish two forms of refinement on the basis of different structural assumptions. These material assumptions also guide us in formulating the adequate conservative constraints, and these will inevitably vary on a case-by-case basis. Our approach stands in stark contrast with much of the literature on awareness growth from a Bayesian perspective. This literature is primarily concerned with a formal, almost algorithmic solution to the problem. We suspect that seeking such formal solution is doomed to fail. Insofar as Reverse Bayesianism is an expression of of this formalistic aspiration, we agree with Steele and Stefánsson that we are better off looking elsewhere.

7 Extra Materials

7.1 Expansion

There remains to examine cases of awareness expansion. They consist in the addition of another proposition not previously in the algebra, but that is not a refinement of existing propositions. The addition of the hypothesis *Surprise* is, however, an ambiguous case. For one thing, *Surprise* is a novel hypothesis that cannot be subsumed under *Friends/acquaintances* or *Affair*. On the other, *Surprise* seems a refinement of *Friends/acquaintances*, since a meeting for planning a surprise is a more specific way to describe a meeting of acquaintances. A more clear-cut case of awareness expansion would be the following. The police is investigating a murder case. There are two suspects under investigation: Joe and Sue. They both have a motive. The incriminating evidence favors one over the other, but not overwhelmingly. T hen, a new hypothesis is considered: Ela could be the perpetrator. The evidence incriminates Ela almost without any doubt. Any theory of awareness growth should be able to model the difference between the example provided by Steele and Stefánsson and the criminal case just outlined.

[8] A generalzied version iof ($C^*) wouldbe as follows. $Let E be a node with several incoming arrows departing from a finite set of nodes, $X_1, X_2, \ldots X_k$ that are shared before and after awareness growth. Then:

$$\mathsf{P}(E = x | X_1 = x_1 \wedge X_2 = x_2 \ldots X_k = x_k) \geq \mathsf{P}(E = x | X_1 = x_1' \wedge X_2 = x_2' \ldots X_k = x_k')$$

iff

$$\mathsf{P}^+(E = x | X_1 = x_1 \wedge X_2 = x_2 \ldots X_k = x_k) \geq \mathsf{P}^+(E = x | X_1 = x_1' \wedge X_2 = x_2' \ldots X_k = x_k').$$

They are both, arguably, cases of expansion, but they are also different.

Steele and Stefánsson provide a formal definition of the difference between refinement and expansion. Our observations here are largely confined at the intuitively level. Our point is that there are a number of intuitively plausible differences that a formal theory should be able to capture. The coarse distinction between refinement and expansion might be, in the end, too coarse. Relying on Bayesian networks, we will illustrate this point more precisely in the next section.

7.2 Steele and Stefánsson example

Before awareness growth, the Bayesian network has a simple form:

$$H \rightarrow E,$$

where the hypothesis variable H takes two values, $H = \textit{Affair}$ and $H = \textit{Friends/acquaintances}$. The evidence variable E can take several values, one of them being $E = \textit{Secretive}$. You could have seen other things other than what you saw, but there is no need to specify the other values exhaustively. Suppose the prior odds ratio of the hypotheses is 1:1, say, because you suspected your partner might be cheating on you, and the likelihood ratio

$$\frac{\mathsf{P}(E = \textit{Secretive}|H = \textit{Affair})}{\mathsf{P}(E = \textit{Secretive}|H = \textit{Friends/acquaintances})}$$

is 9:1, because the hypothesis *Affair* is a better explanation of the evidence than the hypothesis *Friends/acquaintances*. Then, the posterior probability given the evidence

$$\mathsf{P}(H = \textit{Affair}|E = \textit{Secretive})$$

is quite high, $\frac{9}{10} = .9$. So $\mathsf{P}^{E=\textit{Secretive}}(H = \textit{Affair}) = .9$.[9]

After awareness growth, the Bayesian network should be modified as follows:

$$H \leftarrow H' \rightarrow E,$$

where the new hypothesis node now consists of three values instead of two:

$H' = \textit{Affair}$

$H' = \textit{Friends/acquaintances} \wedge \neg \textit{Surprise}$

$H' = \textit{Friends/acquaintances} \wedge \textit{Surprise}$.

The scenario *Friends/acquaintances* is split into the scenario in which your partner and best friend met simply as friends or acquaintances, and the scenario in which they met to prepare a surprise party for you. On this interpretation, the counterexample by Steele and Stefánsson is a case of refinement, not expansion. We will return to this point later.

The network contains a directed arrow between the old hypothesis node H and the new hypothesis node H' This arrow can be interpreted as a bridge between the old awareness state limited to two hypotheses and the new awareness state that contains an additional hypothesis. This bridge is purely conceptual and can be defined by two sets of constrains. The first set of constrains posits that *Affair* under H has the same meaning as *Affair* under H':

$\mathsf{P}^+(H = \textit{Affair}|H' = \textit{Affair}) = 1$

$\mathsf{P}^+(H = \textit{Affair}|H' = \textit{Friends/acquaintances}) = 0$

$\mathsf{P}^+(H = \textit{Affair}|H' = \textit{Surprise}) = 0$

[9] This calculation presupposes that the two hypotheses *Affair* and *Friends/acquaintances* are exclusive and exhaustive. This assumption is justified given the initial awareness state of the agent.

The second set of constrains posits that hypothesis *Friends/acquaintances* under H can be actually be interpreted in two ways under H', as *Friends/acquaintances* $\wedge \neg$*Surprise* and *Friends/acquaintances* $\wedge$ *Surprise*. So, in other words, the episode of awareness growth consists in the realization that *Friends/acquaintances* can be made precise in two more specific ways:

$\mathsf{P}^{+}(H = \textit{Friends/acquaintances} | H' = \textit{Affair}) = 0$

$\mathsf{P}^{+}(H = \textit{Friends/acquaintances} | H' = \textit{Friends/acquaintances} \wedge \neg \textit{Surprise}) = 1$

$\mathsf{P}^{+}(H = \textit{Friends/acquaintances} | H' = \textit{Friends/acquaintances} \wedge \textit{Surprise}) = 0$

This bridge between H and H' justifies the following conservativity constraint:

$$\frac{\mathsf{P}(E = \textit{Secretive} | H = \textit{Affair})}{\mathsf{P}(E = \textit{Secretive} | H = \textit{Friends/acquaintances})} = \frac{\mathsf{P}^{+}(E = \textit{Secretive} | H = \textit{Affair})}{\mathsf{P}^{+}(E = \textit{Secretive} | H = \textit{Friends/acquaintances})} = \frac{9}{1}$$

7.3 Expansion: criminal case example