Beyond Reverse Bayesianism:

Awareness Growth and Bayesian Networks

Marcello Di Bello and Rafal Urbaniak

June 13, 2022

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).

Both E and H belong to the agent's algebra of propositions. This algebra models the agent's awareness state, the propositions taken to be live possibilities. Conditionalization never modifies the algebra and thus makes it impossible for an agent to learn something they have never thought about. This forces a great deal of rigidity on the learning process. Even before learning about E, the agent must already have assigned a degree of belief to 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.

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).

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. To be sure, the algebra of propositions need not be so narrowly construed that it only contains propositions that are presently under consideration. The algebra may also contain propositions about which, even though they are not the object to present consideration, the agent has already formed, perhaps simplicity, a certain disposition to believe. If, however, we have actual applications of probabilistic tools in mind, this is not a promising strategy. We are not God-like agents, probabilistic models are small-world models always restricted to a pre-specified set of variables, and some guidance as to how these should be revised when our awareness changes without the unrealistic assumption of us already having had the right algebra and probabilities to start with is desirable.

A proposal that has attracted considerable scholarly attention in recent years 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 proportions of probabilities of the propositions shared between the old and new algebra remain the same in the sense to be specified.

Let $\mathscr{F}$ be the initial algebra of propositions and let $\mathscr{F}^+$ the algebra after the agent's awareness state has grown. Both algebras contain the contradictory and tautologous propositions $\bot$ and $\top$, 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 that contain only basic propositions, namely those without 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.

Reverse Bayesianism is an elegant theory that manages to cope with a seemingly intractable problem. As the awareness state 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.

Unfortunately, Reverse Bayesianism 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 share Steele and Stefánsson's skepticism toward Reverse Bayesianism, but also believe their counterexamples have limited applicability. We strengthen their argument by providing a simpler counterexample that is less prone to objections (§ 2) . At the same time, we conjecture that the problem of awareness growth cannot be tackled in an algorithmic manner because subject-matter assumptions, both probabilistic and structural, need to be made explicit. Thanks to its ability to model probabilistic dependencies, we think that the theory of Bayesian networks can help to theorize about awareness growth in the Bayesian framework. We offer two illustrations of this claim. First, we provide an example of awareness growth refinement that it is structurally different from other cases of refinement (§ **??**). Second, we model two scenarios from Anna Mathani, both intended to challenge Reverse Bayesianism (§ 4). As we will see, Bayesian networks allow us to see more clearly which probability assignments should be retained during awareness growth and which ones should be modified. The choice is guided by the underlying structure of the scenarios, requires material knowledge and does not fall out from purely formal constraints.

2 Counterexamples

We begin by rehearsing two of the ingenious counterexamples to Reverse Bayesianism by Steele and Stefánsson. One example targets awareness expansion and the other awareness refinement. The difference between expansion and refinement 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 that the existing partition was not exhaustive.

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)

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, that is, the fact that your partner and your best friend met one night without telling you (*Secretive*). Given this evidence, *Affair* is more probable than *Friends/acquaintances*:

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

When 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*). Given the same evidence, *Friends/acquaintances* is now more likely than *Affair*:

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

This holds assuming that hypothesis *Surprise* is more likely than the hypothesis *Affair*:

$$\mathsf{P}^{+}(\mathit{Surprise}|\mathit{Secretive}) > \mathsf{P}^{+}(\mathit{Affair}|\mathit{Secretive}),$$

and, in addition, that *Surprise* implies *Friends/acquaintances*. After all, in order to prepare a surprise party, your partner and best friend have to be at least acquaintances.

The conjunction of ($>$) and ($<$) violates Reverse Bayesianism. But, as Steele and Stefánsson admits, Reverse Bayesianism 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, the entire possibility space before the episode of awareness growth. In our running example, the proposition $\neg$*Surprise* picks out the entire possibility space in just this way. And conditional on $\neg$*Surprise*, the probability of *Affair* does not change. Thus,

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

Awareness Rigidity is satisfied. 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 that also works against Awareness Rigidity, this time targeting a case of refinement:

> 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. So you assigned the same probability to the hypotheses *Thriller* and *Comedy*. But learning that the owner is simple-minded made you think that the level of linguistic difficulty must be low and the movie most likely a thriller rather than a comedy (perhaps because thrillers are simpler—linguistically—than comedies). So, against Reverse Bayesianism, MOVIES violates the condition $\frac{\mathsf{P}(\mathit{Thriller})}{\mathsf{P}(\mathit{Comedy})} = \frac{\mathsf{P}^{+}(\mathit{Thriller})}{\mathsf{P}^{+}(\mathit{Comedy})}$.

The counterexample also works against Awareness Rigidity. It is not true that $\mathsf{P}(\mathit{Thriller}) = \mathsf{P}^{+}(\mathit{Thriller}|\mathit{Thriller} \vee \mathit{Comedy})$. 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. So the proposition which picks out the entire possibility space should be the same before and after awareness growth, for example, $\mathit{Thriller} \vee \mathit{Comedy}$. In cases of awareness growth by refinement, then, Awareness Rigidity mandates that all probability assignments stay the same. But MOVIES does not satisfy this requirement.

This is all well and good, but how strong of a counterexample is this? Steele and Stefánsson consider an objection:

> It might be argued that our examples are not illustrative of . . . a simple growth in awareness; rather, our examples illustrate and should be expressed formally as complex learning experiences, where first there is a growth in awareness, and then there is a further learning event ... In this way, one could argue that the awareness-growth aspect of the learning event always satisfies Reverse Bayesianism.

Admittedly, MOVIES 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. Could Reserve Bayesianism still work for the first episode, but not the second? Steele and Stefánsson do not address this question explicitly, but insist that no matter the answer both episodes are instances of awareness growth. We agree with them on this point. Awareness growth is both *entertaining* a new proposition not in the initial awareness state of the agent and *learning* a new proposition. Nonetheless, we should still wonder. Is the second episode (learning something new) necessary for the counterexample to work together with the first episode (mere refinement without learning)?

Suppose the counterexample did work only in tandem with an episode of learning something new. If 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 featuring 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.

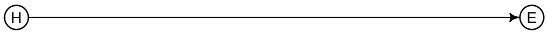
Or perhaps there is a more straightforward counterexample that only depicts mere refinement without an episode of learning and that still challenges Reverse Bayesianism and Awareness Rigidity. This is indeed the case, as this scenario illustrates:

> 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 lower 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.[1] 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 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). Bayesian networks are relied upon in many fields, but have been little used to model awareness growth. We think instead they are a good framework for this purpose. Awareness growth can be modeled as 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 LIGHTING with Bayesian networks, we start with this graph, which is the usual hypothesis-evidence idiom:



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.[2] 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:

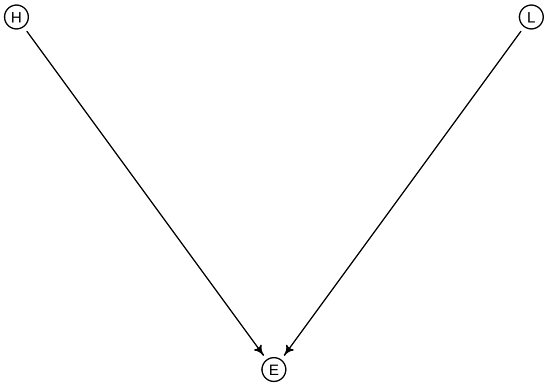
$$\mathsf{P}(E{=}\mathit{seen}|H{=}\mathit{present}) > \mathsf{P}(E{=}\mathit{seen}|H{=}\mathit{absent})$$

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*.

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

[1] Strictly speaking, you are learning that it is *possible* that the lighting conditions were bad. However, you do not condition on the proposition 'the lighting conditions were bad' or 'the lighting conditions were good' as if you learned it with certainty, and thus you do not learn about the lighting conditions in the sense in which learning is understood in this paper.

[2] 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?



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.

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? 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:

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

In our example, many plausible possible probability assignments violate this equality. But it would be quite a coincidence if (C) were true. If before awareness growth you thought the evidence favored the hypothesis H=*present* moderately strongly, after the growth in awareness, the evidence is likely to appear less strong.[3] So, mere refinement can weaken the evidence,

Need a more general argument here. Simulation?

[3] By the law of total probability, the right hand side of the equality should be expanded, as follows:

$$\frac{\mathsf{P}^+(E=e|H=h)}{\mathsf{P}^+(E=e|H=h')} = \frac{\mathsf{P}^+(E{=}seen \wedge L{=}good|H{=}present) + \mathsf{P}^+(E{=}seen \wedge L{=}bad|H{=}present)}{\mathsf{P}^+(E{=}seen \wedge L{=}good|H{=}absent) + \mathsf{P}^+(E{=}seen \wedge L{=}bad|H{=}absent)}.$$

For concreteness, let's use some numbers:

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

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

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

So the ratio $\frac{\mathsf{P}(E{=}seen|H{=}present)}{\mathsf{P}(E{=}seen|H{=}absent)}$ equals 2. After the growth in awareness, the ratio $\frac{\mathsf{P}^+(E{=}seen|H{=}present)}{\mathsf{P}^+(E{=}seen|H{=}absent)}$ will drop to $\frac{.65}{.45} \approx 1.44$. The calculations here rely on the dependency structure encoded in the Bayesian network (see starred step below).

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

even without learning anything new.

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. 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:

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

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*.

3 The importance of structural assumptions

Unlike MOVIES, the counterexample LIGHTING works even though it only depicts a case of awareness growth that consists in refinement without learning. Defenders of Reverse Bayesianism and Awareness Rigidity can no longer claim that their theories work when awareness growth is not intertwined with learning. So, Steele and Stefánsson's critique of these theories sits now on a firmer ground. But there is a more general lesson to be learned here. It has to do with the importance of modeling structural assumptions and the role of Bayesian networks in theorizing about awareness growth.

Steele and Stefánsson's argument relies on the distinction between cases of awareness and cases of expansion. Both MOVIES and LIGHTING are cases of refinement, and they both violate Reverse Bayesianism. FRIENDS, instead, is a case of expansion and does not violate Reverse Bayesian (understood as Awareness Rigidity). But this categorization is too simple. As we shall see, not all cases of refinement are the same, and it is important to understand and to be able to model the structural differences that may arise between them. To illustrate what is at issue, consider this variation of the LIGHTING scenario:

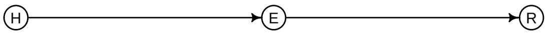
> VERACITY: A witness saw that the defendant was around the crime scene and you initially took this to be evidence that the witness was actually there. But then you worry that the witness might be lying or misremembering what happened. Perhaps, the witness was never there, made things up or mixed things 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 prove to be a good analytic tool here.

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

This argument can be repeated with several other numerical assignments.

The graphical network should initially look like the initial DAG for Lighting. But, as your awareness grows, the graphical network should be updated:



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.

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. In the Bayesian network, this falls out from its structure, as the connection between H and E remains unchanged. Thus, Reverse Bayesianism and Awareness Rigidity are perfectly fine in scenarios such as Veracity.

A possible 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 belief in what they said about the defendant’s whereabouts? Surely so. But note that in Veracity an episode of awareness refinement takes 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 the growth in awareness, you realize that 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* conditional on your evidence, but this does not conflict with Reverse Bayesianism or Awareness Rigidity, as you re-conceptualized what your evidence actually is. 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 at least some cases of awareness expansion; (b) it also handles successfully at least some cases cases of refinement such as Veracity; but (c) it does fail in cases of refinement like Lighting. So, ultimately, Steele and Stefánsson’s critique only targets a subclass of refinement cases. The scope of this critique is therefore somewhat limited. And yet, we do not think the prospects for Reverse Bayesianism are good. In this respect, we tend to agree with Steele and Stefánsson. But we conjecture that there is a deeper difficulty for Reverse Bayesianism, besides possible counterexamples that may be leveled against it. The deeper difficulty is that it seeks to provide a formal, almost algorithmic solution to the problem of awareness growth, and this formal aspiration is likely to lead us down the wrong path.

To see why, consider again the distinction between the two cases of refinement. Reverse Bayesianism is perfectly fine in scenarios like Veracity, but fails in scenarios like Lighting. Why is that so? In one scenario, 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. The two scenarios are structurally different, and this difference can be appreciated by looking at the Bayesian networks used to model them. In Veracity, the new node is added downstream. Since the conditional probabilities associated with the upstream nodes are unaffected, Reverse Bayesianism is satisfied. By contrast, in Lighting, the new node is added upstream. Since the conditional probabilities that are associated with the downstream nodes will often have to change, Reverse Bayesianism fails here.

This discussion suggests a conjecture: structural—possibly causal—features about how we conceptualize a specific scenario seems to be the guiding principles about how we update the probability function through awareness growth, not a formal principle like Reverse Bayesianism. We further elaborate on this conjecture in the final section by drawing on some examples from Anna Mathani.

4 Mathani's examples

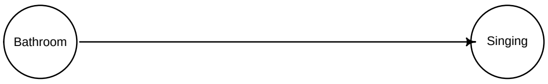
Mahtani offered the following counterexamples:

> TENANT: You are staying at Bob's flat which he shares with his landlord. You know that Bob is a tenant, and that there is only one landlord, and that this landlord also lives in the flat. In the morning you hear singing coming from the shower room, and you try to work out from the sounds who the singer could be. At this point you have two relevant propositions that you consider possible, that it's the landlord singing (*Landlord*), or that Bob is the singer (*Bob*) . . . the possibility suddenly occurs to you that there might be another tenant living in the same flat, and that perhaps that is the person singing in the shower (*Other*).
>
> TENANT: You know that I am holding a fair ten pence coin which I am about to toss. You have a credence of 0.5 that it will land *Heads*, and a credence of 0.5 that it will land *Tails*. You think that the tails side always shows an engraving of a lion. So you also have a credence of 0.5 that it will land with the lion engraving face-up (*Lion*). You become aware that there are some ten pence coins that have an engraving of Stonehenge on the tails side, so you consider the proposition that the toss will end up with Stonehenge face-up (*Stonegenge*).

The problem for Reverse Bayesianism with TENANT appears when you consider the proposition is that the singer is a tenant (*Tenant*). Suppose you originally took $\mathsf{P}(\mathit{Landlord}) = \mathsf{P}(\mathit{Bob}) = .5$. Since *Bob* and *Tenant* originally are the same fact, $\mathsf{P}(\mathit{Tenant})$ = .5. Since these are propositions that you were originally aware of, they should remain in the same proportion after your awareness grows, that is you should have $\mathsf{P}^{+}(\mathit{Landlord}) = \mathsf{P}^{+}(\mathit{Tenant}) = k$. But also *Bob* and *Tenant*, being the same fact, had the same probability, so you should also have $\mathsf{P}^{+}(\mathit{Bob}) = \mathsf{P}^{+}(\mathit{Tenant}) = k$. But now, *Other* entails *Tenant* and *Bob* and *Other* is disjoint, so it follows that $\mathsf{P}^{+}(\mathit{Other}) = 0$, clearly an undesired consequence.

The problem for Reverse Bayesianism with TAILS is that you initially gave *Tails* and *Lion* the same credence so you should have $\mathsf{P}^{+}(\mathit{Tails}) = \mathsf{P}^{+}(\mathit{Lion}) = k$. But since *Lion* and *Stonehenge* are incompatible and the latter entails *Tails*, you should have $\mathsf{P}^{+}(\mathit{Stonehenge})$ = 0$, again an undesirable conclusion.

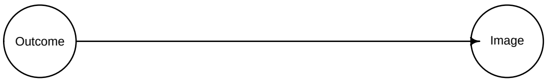
How do we approach these two cases using Bayesian networks? Let's start with TENANTS. We start with the following DAG:



Initially, *Bathroom* has three possible states, representing who is in the bathroom singing: *landlord*, *bob*, and *noone*. *Singing* is a binary node with two possible values: *true*, and *false*. The original example is under-specified, so let's make some probabilities and run with them. Suppose that both Bob and you have the same probability of singing in the bathroom, say .2. Also, quite naturally, the probability of someone singing in an empty bathroom is 0. Say also that the prior probability of you being in the bathroom is .1, the same as the probability of the landlord being in the bathroom. Now you learn $\mathit{Singing} = \mathit{true}$ and update accordingly. The posterior probabilities of *bob* and *landlord* now both equal .5, in line with Mahtani's example. After your awarness grows, *Bathroom* now have one more possible state, *other*, which is also given prior probability of .1 (and the probability of *noone* drops from .8 to .7). Now if you learn *Singing* = *true* and update, the posterior probabilities of *bob*, *tenat* and *other* all equal 1/3, which is exactly as desired.

There are two points here to observe. First, there is no magic in modeling the problem with Bayesian networks. If somehow you know the priors for the candidates are the same (which the example, we take, already assumes), and assume the probabilities that Bob would sing or that the landlord would sing in bathroom is not impacted by the new possibility, the expected outcome falls out. Second, the assumptions made and needed are non-trivial material assumptions that should not and cannot fall out from purely formal considerations. As for priors, we can easily imagine that landlord uses the bathroom in the mornings more often, or tends to sing less. As for the impact of the new tenant, we can also easily imagine circumstances in which Bob is less likely to sing, say because now he is shy to sing in the presence of a new flatmate.

The same two points can be made with respect to TAILS. Now the structure of the scenario is represented by the following DAG:



Outcome has two states, *tails* and *heads*. Initially, image has two states too: *none* and *lions*. We know that images have no impact on the coin's fairness, so the priors for Outcome are .5 each, and this doesn't change once you learn something about other images on the coin (again, a material assumption!). Initially, the probability of *none* given *heads* is 1 and *none* given *tails* is 0. Then you learn that not all images are those of a lion, so you need some new probabilities (not given in the original example). Say *lion* given *tails* has now the probability of .9. What was your original prior on *lion*? .5. What is it after the awareness growth? .45. Again, no surprises in the construction, and again, how we should build the network and which probabilities should shift is based on our material knowledge that does not and should not fall out of purely formal considerations.

5 Conclusion

We argued that Steele and Stefánsson's case against Reverse Bayesianism is weaker than it might seem at first. The scenario MOVIES—which is their key counterexample—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. At the same time, we showed that there are cases of refinement like VERACITY in which Reverse Bayesianism and Awareness Rigidity are perfectly fine in their place. ADD SOME BIT ABOUT HOW WE MODEL MATHANI EXAMPLES

We conclude with a general moral. 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 this formalistic aspiration, we agree with Steele and Stefánsson that we are better off looking elsewhere.