Beyond Reverse Bayesianism:

Awareness Growth in Bayesian Networks

Marcello Di Bello and Rafal Urbaniak

July 28, 2022

We examine Steele and Stefánsson's case against Reverse Bayesianism, a popular theory that addresses the problem of awareness growth. We show that Steele and Stefánsson's counterexamples have limited applicability but agree with their skepticism toward Reverse Bayesianism. We strengthen their argument by providing a simpler counterexample that is less prone to objections. In addition, we submit that the problem of awareness growth cannot be tackled in an algorithmic manner, because subject-matter assumptions needs to be made explicit. Thanks to their ability to express probabilistic dependencies, we sketch how Bayesian networks can help to model awareness growth in the Bayesian framework.

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. 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), as though learning was confined to an 'initial prison' (Lakatos, 1968).

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.

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 which, though outside the agent's present consideration, are still the target, perhaps implicitly, of the

agent's disposition to believe.[1] But even this expanded algebra will have to be revised sooner or later. The algebra of propositions could in principle contain anything that could possibly be conceived, expressed, thought of. Such rich algebra would not need to change at any point, but this is hardly a plausible model of ordinary agents with bounded resources such as ourselves.

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 (Chihara, 1987; Earman, 1992; Glymour, 1980), language changes and paradigm shifts (Williamson, 2003), and theories of induction (Zabell, 1992). A proposal that has attracted considerable scholarly attention in recent years is Reverse Bayesianism (Bradley, 2017; Karni & Vierø, 2015; Wenmackers & Romeijn, 2016). 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 a 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 previously. 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 Bayesian conditionalization which preserves the old probability distribution conditional on what is learned.

Unfortunately, Reverse Bayesianism is not without complications. Steele & Stefánsson (2021) argue that it faces insurmountable difficulties. Their argument rests on a number of ingenious counterexamples. We share Steele and Stefánsson's skepticism toward Reverse Bayesianism, but also believe that their counterexamples have limited applicability (§ 2.1). To remedy that, we provide a simpler counterexample that is less prone to objections (§ 2.2).

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 express probabilistic dependencies, we think that the theory of Bayesian networks can help to model awareness growth in the Bayesian framework. We offer two illustrations of this claim (§ 3). As we will show, 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.

[1] Roussos (2021) notes that, for the sake of clarity, the problem of awareness growth should only address propositions which agents are *truly* unaware of (say new scientific theories), not propositions that were temporarily forgotten or set aside. This is a helpful clarification to keep in mind, although the recent literature on the topic does not make a sharp a distinction between true unawareness and temporary unawareness.

2 Counterexamples

In this section, we rehearse two of the counterexamples to Reverse Bayesianism by Steele and Stefánsson. One example targets awareness expansion and the other awareness refinement (more on this distinction soon). We show why they make a limited case against Reverse Bayesianism and then provide a better counterexample with the aid of Bayesian networks.

2.1 Friends and Movies

The difference between expansion and refinement is intuitively plausible, but 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 occurs when the new proposition is inconsistent with the existing ones, thus making the old partition no longer 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 & Stefánsson, 2021)(Sec. 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*:

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

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*:

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

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

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

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 since *Friends/acquaintances* and *Affair* are basic propositions that do not contain any connectives. 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,

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

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 & Stefánsson, 2021)(Sec. 5, Example 3)

This is a case of refinement. For you initially categorized movies by just language and genre, and then you refined your categorization by adding another variable, level of difficulty. Without considering language difficulty, you assigned the same probability to the hypotheses *Thriller* and *Comedy*. But learning that the owner was 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 $\frac{\mathsf{P}(\textit{Thriller})}{\mathsf{P}(\textit{Comedy})} = \frac{\mathsf{P}^+(\textit{Thriller})}{\mathsf{P}^+(\textit{Comedy})}$.

The counterexample also violates Awareness Rigidity. For consider a proposition that picks out the entire possibility space, for example, $\textit{Thriller} \vee \textit{Comedy}$.[2] Awareness Rigidity would require that $\mathsf{P}(\textit{Thriller}) = \mathsf{P}^+(\textit{Thriller}|\textit{Thriller} \vee \textit{Comedy})$. But MOVIES does not satisfy this equality since the probability of *Thriller* has gone up.

How good 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, many could still wonder. Is the second episode (learning something new) necessary for the counterexample to work together with the first episode (mere refinement without learning)?

[2] Since MOVIES is a case of refinement, $\textit{Thriller} \vee \textit{Comedy}$ picks out the entire possibility space both before and after awareness growth.

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.

2.2 Lighting

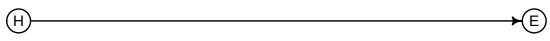
We claim there is a more straightforward counterexample that only depicts mere refinement without an episode of learning and that still challenges Reverse Bayesianism and Awareness Rigidity. To see that this is indeed the case, we propose to consider the following scenario:

> 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.[3] Still, mere refinement in this scenario challenges Reverse Bayesianism and Awareness Rigidity. That this should be so is not easy to see. We rely on the theory of Bayesian networks to see why.

A Bayesian network is compact formalism to represent probabilistic dependencies. A Bayesian network consists of a direct acyclic graph (DAG) 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. Bayesian networks are relied upon in many fields, but have been rarely deployed to model awareness growth (the exception is Williamson (2003)). 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 the prior probabilities for all the states of H, and conditional probabilities of the form $\mathsf{P}(E = e|H = h)$, where uppercase letters represent the variables (nodes) and lower case letters represent the

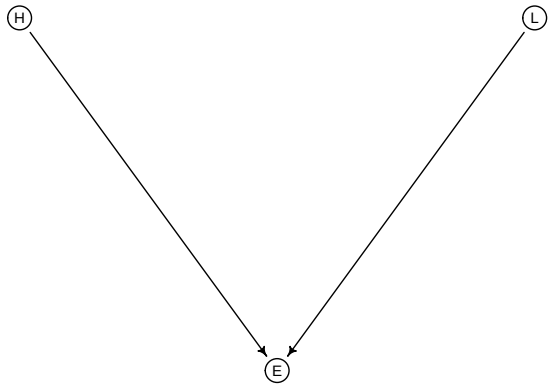
[3] The process of awareness growth in LIGHTING adds only one extra variable, lighting conditions, while MOVIES adds two extra variables, language difficulty and whether the owner is simple-minded or not. Further, MOVIES contains a clear-cut case of learning, that is owner *is* simple-minded. This is not so in in LIGHTING. Strictly speaking, you are learning that it is *possible* that the lighting conditions were bad. However, you are not conditioning on the proposition 'the lighting conditions were bad' or 'the lighting conditions were good.' So you are not learning about the lighting conditions in the sense in which learning is understood in this paper.

values of these variables.[4] Since you trust the evidence, you think that the evidence is more likely under the hypothesis that the defendant was present at the crime scene than under the alternative hypothesis:

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

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:



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:

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

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

Here is what you are thinking: if the lighting conditions were good, you should still trust the evidence like you did before (first line). But if the lighting conditions were bad, you should regard the evidence as no better than chance (second line).

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 changing the probability function from $\mathsf{P}()$ to $\mathsf{P}^{+}()$, it would be quite a coincidence if (C) were true. In our example, many possible probability assignments violate this equality. If before awareness growth you thought the evidence favored the hypothesis *H=present* to some extent, after the growth in awareness, the evidence is likely to appear less strong.[5]

[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; see footnote 11.

[5] By the law of total probability, the right hand side of the equality in (C) should be expanded, as follows:

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

Why does this matter? We have seen that, after awareness growth, you should typically regard the evidence *E=seen* 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 which, in cases of refinement, requires that the probability of basic propositions stay fixed.[6]

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*.

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.

For concreteness, let's use some numbers:

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

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

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

$$\mathsf{P}^{+}(L{=}bad) = \mathsf{P}^{+}(L{=}good) = .5.$$

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).

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

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

This argument can be repeated with many other numerical assignments.

[6] Awareness Rigidity requires that $\mathsf{P}^{+}(A|T^{*}) = \mathsf{P}(A)$, 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. But, in cases of refinement, T^{*} does not change, so in such cases, Awareness Rigidity requires $\mathsf{P}^{+}(A) = \mathsf{P}(A)$.

3 Structural assumptions

Besides counterexamples that can be leveled against Reverse Bayesianism, we think there is a more general lesson to be learned. It has to do with the importance of formalizing structural assumptions and the role of Bayesian networks in modeling awareness growth. We substantiate this point with two illustrations. The first shows that the distinction between refinement and expansion that Steele and Stefánsson rely on is more fine-grained. The second illustration draws on some scenarios by Anna Mathani (2020).

3.1 Another refinement

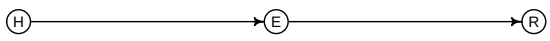
Steele and Stefánsson's argument relies on the distinction between refinement and expansion. But this categorization is too simple. As we shall see, not all cases of refinement are the same, and it is important to understand and to be able to model the structural differences that may arise between them. For 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.

Should you reassess the evidence at your disposal? If so, how? At first, it might seem that this scenario is no different from LIGHTING. The realization that lighting could be bad should make you less confident in the truthfulness of the sensory evidence. And the same conclusion should presumably follow from the realization that the witness could be lying.

But, upon closer scrutiny, things are not as simple as they might seem at first. The evidence at your disposal in LIGHTING is the sensory evidence—the experience of seeing—and the possibility of bad lighting does warrant lowering your confidence in the truthfulness of the visual experience. But the possibility of lying in VERACITY does not warrant lowering your confidence in the truthfulness of the visual experience, only in the truthfulness of the *reporting* of the experience. The distinction between the visual experience and its reporting is crucial here. Bayesian networks help to model this distinction precisely, and in this way see why LIGHTING and VERACITY are structurally different cases of refinement.

The graphical network should initially look like the initial DAG for LIGHTING, consisting of the hypothesis node H upstream and the evidence node E downstream. As your awareness grows, the graphical network should be updated by adding another node R further downstream:



As before, the hypothesis node H bears on the whereabouts of the defendant and has two values, *H=present* and *H=absent*. Note the difference between E and R. The evidence node E bears on the visual experience had by the witness. The reporting node R, instead, bears on what the witness reports to have seen. The chain of transmission from 'visual experience' to 'reporting' may fail for various reasons, such as lying or misremembering.

In VERACITY, 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 comparing the old and new Bayesian network, this equality falls out from their structure, as the connection between H and E remains unchanged. Thus, Reverse Bayesianism and Awareness Rigidity are perfectly fine in scenarios such as VERACITY. This does not mean that the assessment of the probability of the hypothesis *H=present* should undergo no change. If you worry that the witness could have lied, this should presumably make you less confident about *H=present*. To accommodate this intuition, VERACITY can

be interpreted as a scenario in which an episode of awareness refinement takes place together with a form of retraction. At first, after the learning episode, you update your belief based on the *visaul experience* of the witness. But after the growth in awareness, you realize that your learning is in fact limited to what the witness *reported* to have seen. The previous learning episode is retracted and replaced by a more careful statement of what you learned: instead of conditioning on *E=seen*, you should condition on what the witness reported to have seen, *R=seen-reported*. This retraction will affect the probability of the hypothesis *H=present*.

Where does this leave us? Refinement cases that might at first appear similar can be structurally different in important ways, and this difference can be appreciated by looking at the Bayesian networks used to model them. In modeling VERACITY, the new node is added downstream, while in modeling LIGHTING, it is added upstream. This difference affects how probability assignments should be revised. Since the conditional probabilities associated with the upstream nodes are unaffected, Reverse Bayesianism is satisfied in VERACITY.[7] By contrast, since the conditional probabilities associated with the downstream node will often have to change, Reverse Bayesianism fails in LIGHTING.

This discussion suggests a conjecture: structural 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 by drawing on some examples from Anna Mathani.

3.2 Mathani's counterexamples

Mathani (2020) offers two counterexamples to Reverse Bayesianism. The first goes like this:

> TENANT: Suppose that 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 ... *Landlord* standing for the possibility that the landlord is the singer, and *Bob* standing for the possibility that Bob is the singer ... Because you know that Bob is a tenant in the flat, you also have a credence in the proposition *Tenant* that the singer is a tenant. Your credence in *Tenant* is the same as your credence in *Bob*, for given your state of awareness these two propositions are equivalent ... Now let's suppose the possibility suddenly occurs to you that there might be another tenant living in the same flat (*Other*).

Initially, you thought the singer could either be the landlord or Bob, the tenant. Then you come to the realization that a third person could be the singer, another tenant. Before awareness growth, that Bob is in the shower and that a tenant is in the shower are equivalent descriptions. After awareness growth, this equivalence breaks down.

Why is this scenario problematic? Suppose, after you hear singing in the shower, you become sure someone is in there, but you cannot tell who. So $\mathsf{P}(\mathit{Landlord}) = \mathsf{P}(\mathit{Bob}) = 1/2$, and since *Bob* and *Tenant* are equivalent, also $\mathsf{P}(\mathit{Tenant}) = 1/2$. Now, *Landlord*, *Bob* and *Tenant* are all propositions that you were originally aware of, and thus Reverse Bayesianisn requires that their assigned probabilities should remain in the same proportion after your awareness grows. But note that *Other* entails *Tenant* and *Bob* and *Other* are disjoint, so it follows that $\mathsf{P}^+(\mathit{Other})$ must have zero probability.[8] This is an undesired outcome that rules out the possibility that

[7] Note that $\mathsf{P}(H{=}\mathit{present}|E{=}\mathit{seen}) \neq \mathsf{P}(H{=}\mathit{present}|R{=}\mathit{seen\text{-}reported})$, but since you are conditioning on different propositions, this does not conflict with Reverse Bayesianism.

[8] If $\mathsf{P}^+(\mathit{Other}) > 0$, the proportion of *Tenant* to *Landlord* or the proportion of *Bob* to *Landlord* should change.

there could be a third person in the shower.[9]

Consider now Mathani's second counterexample:

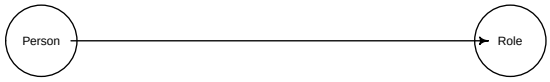
> COIN: You know that I am holding a fair ten pence UK coin which I am about to toss. You have a credence of 0.5 that it will land *Heads*, and a credence of 0.5 that it will land *Tails*. You think that the tails side always shows an engraving of a lion. So you also have a credence of 0.5 that (*Lion*) it will land with the lion engraving face-up: relative to your state of awareness *Tails* and *Lion* are equivalent.... Now let's suppose that you somehow become aware that occasionally ten pence coins have .... an engraving of Stonehenge on the tails side.

Tails and *Lion* are equivalent propositions prior to awareness growth. Suppose you initially gave *Tails* and *Lion* the same credence. Reverse Bayesianism requires that their relative proportions should stay the same after awareness grow. The same applies to *Heads* and *Tails*. But since *Lion* and *Stonehenge* are incompatible and the latter entails *Tails*, you should have $\mathsf{P}^{+}(Stonehenge) = 0$, again an undesirable conclusion.

Mathani notes that COIN has the same structure as TENANT. This is true to some extent, but there is also an interesting asymmetry between the two scenarios. In TENANT, it is natural to assign $1/3$ to *Landlord*, *Bob* and *Other* after awareness growth. That someone is singing in the shower is evidence that someone must be in there, but without any more discriminating evidence, each person should be assigned the same probability. Consequently, a probability of 2/3 should be assigned to *Tenant*. On this picture, the proportion of *Landlord* to *Tenant* changes from 1:1 (before awareness growth) to 1:2 (after awareness growth). But, in COIN, the relative proportion of *Heads* to *Tails* should remain constant throughout, unless evidence emerges that the coin is not fair. One might have expected that *Landlord* and *Tenant* would behave just like *Heads* and *Tails*, but actually they do not.

Bayesian networks can help to model the asymmetry between these two scenarios. Consider COIN first. The structure of the scenario is represented by the following graph:

The upstream node *Outcome* has two states, *tails* and *heads*. These two states remain the same throughout. What changes are the states associated with the *Imagine* node downstream. Before awareness growth, the node *Image* has two states: *lions* and *heads-image*.[10] You assume that *Image* = *lions* is true if and only if *Outcome* = *tails* is true. Then, you come to the realization that the imagines for tails include a lion or a stonehenge engraving. So, after awareness growth, the node *Image* contains three states: *lion*, *stonehenge* and *heads-image*. Consider now the other scenario, TENANT. We start with the following graph:



Initially, the upstream node *Person* has two possible states, representing who is in the bathroom singing: *landlord-person* and *bob*. To simplify things, the assumption here is that the evidence of singing has already ruled out the possibility that no one would be in the shower. The downstream node *Role* has also two values, *landlord* and *tenant*. After your awareness grows, the upstream node *Person* should now have one more possible state, *other*.

[9] Awareness Rigidity is no of help either because it would require that $\mathsf{P}^{+}(Landlord|Landlord \vee Tenant) = \mathsf{P}^{+}(Bob|Landlord \vee Tenant)$ both equal $1/2$, thus forcing $\mathsf{P}^{+}(Other|Landlord \vee Tenant)$ to zero.

[10] The heads side must have some image, not specified in the scenario.

The difference in modeling the two scenarios is this. In COIN, the states of the upstream node remain fixed, whereas in TENANT, they change. After awareness growth, no new state is added to *Outcome*, but an additional state, *other*, is added to *Person*. Plausible probability distributions for the Bayesian networks associated with the two scenarios are displayed in Table 1. How the networks should be built and which probabilities should shift is based on our background knowledge. This knowledge tells us that the equiprobability of *heads* and *tails* should not be affected by realizing that *stonhenge* is another possible engraving for the tails side. It also tells us that the probabilities of *landlord* and *tenant* should be affected by realizing that a third person could be in the shower.

We conclude with some programmatic remarks. We think that the awareness of agents grows while holding fixed certain material structural assumptions, based on commonsense, semantic stipulations or causal dependency.[11] 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. These material assumptions also guide us in formulating the adequate conservative constraints, and these will inevitably vary on a case-by-case basis. The literature on awareness growth from a Bayesian perspective is primarily concerned with a formal, almost algorithmic solution to the problem. Insofar as Reverse Bayesianism is an expression of this formalistic aspiration, we agree with Steele and Stefánsson that we are better off looking elsewhere.

References

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.

Chihara, C. S. (1987). Some problems for bayesian confirmation theory. *British Journal for the Philosophy of Science*, *38*(4), 551–560.

Earman, J. (1992). *Bayes or bust? A critical examination of bayesian confirmation theory*. MIT press.

Glymour, C. (1980). *Theory and evidence*. Princeton University Press.

Howson, C. (1976). The development of logical probability. In *Essays in memory of imre lakatos. Boston studies in the philosophy of science* (pp. 277–298). Springer.

Karni, E., & Vierø, M.-L. (2015). Probabilistic sophistication and reverse bayesianism. *Journal of Risk and Uncertainty Volume*, *50*, 189–208.

Lakatos, I. (1968). Changes in the problem of inductive logic. *Studies in Logic and the Foundations of Mathematics*, *51*, 315–417.

Mathani, A. (2020). Awareness growth and dispositional attitudes. *Synthese*, *198*(9), 8981–8997.

Roussos, J. (2021). Awareness growth and belief revision. *Manuscript*.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical Studies*, *173*, 49–100.

Steele, K., & Stefánsson, O. (2021). Belief revision for growing awareness. *Mind*, *130*(520), 1207–1232.

Wenmackers, S., & Romeijn, J.-W. (2016). New theory about old evidence: A framework for open-minded bayesianism. *Synthese Volume*, *193*, 1225–1250.

Williamson, J. (2003). Bayesianism and language change. *Journal of Logic, Language, and Information*, *12*(1), 53–97.

Zabell, S. (1992). Predicting the unpredictable. *Synthese*, *90*(1), 205–232.

[11] Arrows in Bayesian networks are often taken to represent causal relationships, but other interpretations exist. Schaffer (2016) discusses an interpretation in which arrows represent grounding relations rather than causality.

$\mathsf{P}(Image\|Outcome)$		$Outcome$	
		heads	*tails*
Image	*lion*	0	1
	heads-image	1	0

$\mathsf{P}^+(Image\|Outcome)$		$Outcome$	
		heads	*tails*
Image	*lion*	0	1/2
	stonehenge	0	1/2
	heads-image	1	0

$\mathsf{P}(Outcome) = \mathsf{P}^+(Outcome)$	$Outcome$	
	heads	*tails*
	1/2	1/2

$\mathsf{P}(Role\|Person)$		$Person$	
		landlord-person	*bob*
Role	*tenant*	0	1
	landlord	1	0

$\mathsf{P}^+(Role\|Person)$		$Person$		
		landlord-person	*bob*	*other*
Role	*tenant*	0	1/2	1/2
	landlord	1	0	0

$\mathsf{P}(Person)$	$Person$	
	landlord-person	*bob*
	1/2	1/2

$\mathsf{P}^+(Person)$	$Person$		
	landlord-person	*bob*	*other*
	1/3	1/3	1/3

Table 1: Top table displays a plausible probability distribution for COIN and bottom table does the same for TENANT.