# 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 $A$ after a learning experience $E$ should be the same as the agent's old degree of belief in $A$ conditional on $E$. That is,

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

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 target of extensive discussion.[1] The other assumption—which we will focus on—is that $E$ and $A$ belong to the agent's algebra of propositions. This algebra models the propositions that the agent is aware of and entertains as live possibilities.

The algebra—the agent's awareness state—is fixed once and for all. The learning experience does not modify it. Crucially, even before learning about $E$, the agent already knows the degree of belief in any proposition conditional on $E$. In this model, the agent cannot learn something they have never thought about. This picture forces a great deal of rigidity on the learning process. It 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.[2]

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.

[2] 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). We will return to this distinction in due course.

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 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. For reason that will soon become clear, let's pick out subsets of these algebras which contain only basic propositions, those that do not contain connectives such as negations, conjunctions or disjunctions. Call these subsets $X$ and $X^+$ respectively. Obviously, $\mathscr{F} \subseteq \mathscr{F}^+$ and $X \subseteq X^+$. Reverse Bayesianism posits that the ratio of probabilities for any propositions $A$ and $B$ in $X$—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 an agent's awareness grows, the agent might not want to throw away completely the epistemic work they have done so far. The agent may prefer to retain as much of their old assignments of degrees of beliefs as possible. Reverse Bayesianism provides a simple recipe to do that. It also coheres with the conservative spirit of conditionalization. Reverse Bayesianism preserves the old probability distribution conditional on the old awareness state.[3] Similarly, conditionalization preserves the old probability distribution conditional on what has been learned. **SHOW PICTURE BELOW**.

Reverse Bayesianism is an elegant theory that manages to cope with a seemingly intractable challenge for Bayesianism. Unfortunately, it is not without difficulties. 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 cases of *awareness expansion*. But they claim it cannot work in cases of *awareness refinement*. The distinction between refinement and expansion that Steele and Stefánsson draw, albeit a good first approximation, is too coarse and should be made more precise. We will show that there are cases of refinement in which Reverse Bayesianism (or, at least, a suitable formulation of it) can be made to work.

More generally, we believe that Steele and Stefánsson's negative conclusion about Reverse Bayesianism is overly pessimistic. Much of the literature on awareness growth is concerned with a formal, algorithmic solution to the problem. We believe that seeking a formal, algorithmic solution is not a promising strategy—and in that we agree with Steele and Stefánsson. At the same time, 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 model these material structural assumptions. We sketch how this can done using Bayesian networks. This approach—we think—salvages the spirit of Reverse Bayesianism.

[3] Strictly speaking, this interpretation is what we later call Awareness Rigidity.

## 2 Counterexamples?

To assess the strength of their case against Reverse Bayesianism, we begin by rehearsing some of the ingenious counterexamples that Steele and Stefánsson have formulated. The first is this:

> 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)

So, 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.

Why does this scenario 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. Clearly, 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.[4]

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*:

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

So far so good. Reverse Bayesianism is not yet in trouble. Steele and Stefánsson, however, concludes that the probability of *Friends/acquaintances* should now exceed that of *Affair* ('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.'):

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

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

[4] 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.

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 by replacing it 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:

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

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*):

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

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):

> 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)

Admittedly, this counterexample is complex. For the sake of clarity, it can be split into two episodes of awareness growth. The first episode consists in considering, as a new variable besides language and genre, the language difficulty of the movie. Initially, the algebra contained the propositions *French* and *German*, as well as *Thriller* and *Comedy*. Then, you realize another variable might be at play, namely the level of difficulty of the language of the movie, *Difficult* and *Easy*. The second episode of awareness growth consisting in learning something you did consider before, namely that the owner is simple-minded.

This counterexample is a case of refinement because, first, you categorize movies by just language and genre, and then you add a further category, level of difficulty.[5] The second episode of awareness growth brings further refinement, namely learning that the owner is simple-minded. The difference between the two episodes of awareness growth is that the first did not bring about any change in the probabilities, only the second did. The realization that

[5] In the other counterexample, instead, the possibility space grew by adding situations in which your partner and best friends met neither as lovers nor solely as friends. More on expansion below.

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 (possibly because thrillers are simpler—linguistically—than comedies).

So, as already noted, the counterexample is complex. But, taken at face value, it challenges both Reverse Bayesianism and Awareness Rigidity. It is not true, against Reverse Bayesianism, that $\frac{\mathsf{P}(\textit{Thriller})}{\mathsf{P}(\textit{Comedy})} = \frac{\mathsf{P}^{+}(\textit{Thriller})}{\mathsf{P}^{+}(\textit{Comedy})}$. Further, against Awareness Rigidity, it is not true that $\mathsf{P}(\textit{Thriller}) = \mathsf{P}^{+}(\textit{Thriller}|\textit{Thriller} \vee \textit{Comedy})$. Since this is a case of refinement, the proposition $\textit{Thriller} \vee \textit{Comedy}$ which picks out the entire possibility space is the the same before and after awareness growth. In cases of awareness growth by refinement, then, Awareness Rigidity mandates that all probability assignments stay the same.

This counterexample is likely to leave many unconvinced, or at best puzzled. Since it consists of two episodes, we are left wondering which one of the two is essential. The first is a case of mere refinement—you simply entertain a new way to categorize movies. The second episode is a case of learning—you learn something your did consider before, namely that the owner is simple-minded. So awareness growth is here understood both as (i) *entertaining* a new proposition not in the initial awareness state of the agent (without yet learning it) and (ii) *learning* a new proposition not in the initial awareness state of the agent. We agree with Steele and Stefánsson that both (i) and (ii) shoud count as instance of awareness of awareness growth. Our qualms are about their role in the counterexample. Is only the second episode necessary for the counterexample to work, while the first just gives added context? What is going on, exactly? Is there a more straightforward counterexample that only depicts mere refinement without an episode of learning intertwined with it? A cleaner picture of awareness growth in cases of refinement is preferable.

The need for a cleaner picture also applies to the first counterexample. There remains—we think—the need to further 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 evidence consists in a DNA match and information about how the crime was committed. Sue genetically matches the traces, but is quite short and the perpetrator is known to be a tall person. Joe is neither tall nor does he genetically match the crime traces. In light of the evidence, Sue seems more likely the culprit than Joe, but matters are still open ended. Then, a new hypothesis is considered: Ela could be the perpetrator. As it turns out, Ela genetically matches the traces, is tall enough to have committed the crime, and does have a motive. This seems a straightforward case of expansion because Ela, Sue and Joe are incompatible hypotheses, while *Friends/acquaintances* and *Surprise* need not be. Any model of awareness growth should be able to analyze more precisely the difference between the example provided by Steele and Stefánsson and the criminal case just outlined. 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 too coarse. Relying on Bayesian networks, we illustrate this point more precisely in the next section.

## 3 Bayesian Networks

The key idea is that an awareness state is represented by a graphical networks, with nodes and arrows, along with a probability distribution associated with the network. Awareness growth brings a bout 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 next network. Call this network refinement. While this process is not merely formal or algorithmic, certain plausible conservative constrains guide network refinement.

ADD BRIEF INFORMAL DESCRIPTION ABOUT BAYESIAN NETWORKS HERE.

### 3.1 A simpler counterexample

As noted earlier, Steele and Stefánsson's counterexample to reverse Bayesianism for the case of refinement is rather complex, perhaps unnecessary so. We now present a much simpler one:

> You have evidence that favors a certain hypothesis, say a witness testifies they 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? You come to the following judgment: if the lighting conditions were good, you should still trust the evidence, but if they were bad, you should not. Suppose you do learn that the lighting conditions were bad. In that case, the evidence at your disposal should no longer favor the hypothesis that the defendant was actually around the crime scene. After your awareness has grown, the probabilty of that hypothesis has gone down.

This scenario of awareness growth consists of two parts: first, an episode of mere refinement. You wonder about the lighting conditions. The witness could have seen the defendant under good or bad lighting conditions. After that realization, or concomitantly to it, you learn that the lighting conditions were actually bad. This two-part structure resembles the more complex scenario by Steele and Stefánsson. It is instructive to make our counterexample even simpler and focus on mere refinement without learning. Does mere refinement in this scenario constitute a counterexample to reverse Bayesian and Awareness Rigidity? It does, and the theory of Bayesian networks helps to see why.

Initially, your graphical network looks like this:

$$H \to E$$

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:

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

It is not necessary to fix exact numerical values for these conditional probabilities. Think of the inequality as as a qualitative ordering of how plausible the evidence at your disposal 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*:

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

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

$$H \to E \leftarrow L,$$

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:

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

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

This is what you are thinking: if the lighting conditions were good, you still trust the evidence like you did before. But if the lighting conditions were bad, you regard the evidence as no better than chance. Again, there are no exact numerical values here.

The probability function changes from $\mathsf{P}()$ to $\mathsf{P}^+()$, but the two agree in one respect:

$$\mathsf{P}(E=x|H=x) \geq \mathsf{P}(E=x|H=x') \text{ } \mathit{iff} \text{ } \mathsf{P}^+(E=x|H=x) \geq \mathsf{P}^+(E=x|H=x'), \qquad \text{(C)}$$

where both $E$ and $H$ are nodes that are part of the graphical network before and after the awareness growth. So the plausibility ordering between hypotheses and evidence is preserved. Condition (C) is a good candidate for a conversativity constraint that governs the relationship between $\mathsf{P}()$ and $\mathsf{P}^+()$ (more on this later). It would be, however, too strong to require:

$$\frac{\mathsf{P}(E=x|H=x)}{\mathsf{P}(E=x|H=x')} = \frac{\mathsf{P}^+(E=x|H=x)}{\mathsf{P}^+(E=x|H=x')}. \qquad \text{(CC)}$$

Our counterexample does not violate (C), but does violate (CC). For suppose:

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

$$\mathsf{P}(E{=}seen|H{=}absent) = \mathsf{P}^+(E{=}seen|H{=}absent \wedge L{=}good) = 0.4$$

$$\mathsf{P}^+(E{=}seen|H{=}present \wedge L{=}bad) = \mathsf{P}^+(E{=}seen|H{=}absent \wedge L{=}bad) = 0.5.$$

So the ratio $\frac{\mathsf{P}(E{=}seen|H{=}present)}{\mathsf{P}(E{=}seen|H{=}absent)} = 2$. Before awareness growth, you think the evidence favor the hypothesis moderly strongly. That seem entirely reasonable. But, after the awareness growth, the ratio $\frac{\mathsf{P}^+(E{=}seen|H{=}present)}{\mathsf{P}^+(E{=}seen|H{=}absent)} = \frac{.65}{.45} \approx 1.44$.[6] This shows that mere refinement can weaken the strength of 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:

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

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

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

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

where $\mathsf{P}^{+,L=bad}()$ is the new probability function after learning that *L=bad*. In any event, after awareness growth, the evidence at your disposal favors *H=defendant-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

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

$$\mathsf{P}^{+,L=bad}(H{=}\textit{defendant-present}|E{=}\textit{defendant-seen}) \neq \mathsf{P}(H{=}\textit{defendant-present}|E{=}\textit{defendant-seen})$$

This outcome violates both Reverse Bayesianism and Awareness Rigidity.**WHY? EXPLAIN**.

So we now have a simpler counterexample that works even without postulating that mere refinement takes place together with learning of something new. At the same time, this counterexample also show that something can be preserved, namely the plausibility ordering along the lines of condition (C). This condition should also hold in Steele and Stefánsson's scenario. We simply sketch the reasoning here.
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 have no reason to think the language or genre or the movie is one of the two. So you are indifferent about the different options. But let's 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 more then everything 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, for all combinations, by conditional probabilities 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 consists in realizing that the linguistic difficulty of the movie could also be a factor. So the expanded graphical network now becomes: **DRWA GRAPH WITH AN EXTRA NODE FOR "DIFFICULTY" WITH ARROW POINT INTO "ENJOYMENT" NODE**
The new node can take two values: *Difficulty=high* and *Difficulty=low*. 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:

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

There may well be cases in which this plausibility ordering is not preserved. This would be *truly transformative* awareness growth. But this would perhaps be extremely rare. Suppose you have evidence that—you think—reliably tracks a hypothesis, say evidence as of hand reliably tracks the presence of hands before you. You may now entertain a skeptical hypothesis in which a demon actually switches things around: when you see a hand, there is no hand at, and when you do not see a hand, there is actually a hand. Even in this case, it is not obvious that there would be a violation of a condition like (C). EXPLAIN.

### 3.2 A different refinement

### 3.3 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, $H = \mathit{Affair}$ and $H = \mathit{Friends/acquaintances}$. The evidence variable $E$ can take several values, one of them being $E = \mathit{Secretive}$. 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

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

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

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

is quite high, $\frac{9}{10} = .9$. So $\mathsf{P}^{E=\mathit{Secretive}}(H = \mathit{Affair}) = .9$.[7]

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:

$H' = \mathit{Affair}$
$H' = \mathit{Friends/acquaintances} \wedge \neg \mathit{Surprise}$
$H' = \mathit{Friends/acquaintances} \wedge \mathit{Surprise}$.

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 $H'$:

$\mathsf{P}^+(H = \mathit{Affair}|H' = \mathit{Affair}) = 1$
$\mathsf{P}^+(H = \mathit{Affair}|H' = \mathit{Friends/acquaintances}) = 0$
$\mathsf{P}^+(H = \mathit{Affair}|H' = \mathit{Surprise}) = 0$

The second set of constrains posits that hypothesis *Friends/acquaintances* under $H$ can be actually be interpreted in two ways under $H'$, as $\mathit{Friends/acquaintances} \wedge \neg \mathit{Surprise}$ and $\mathit{Friends/acquaintances} \wedge \mathit{Surprise}$. 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:

$\mathsf{P}^+(H = \mathit{Friends/acquaintances}|H' = \mathit{Affair}) = 0$
$\mathsf{P}^+(H = \mathit{Friends/acquaintances}|H' = \mathit{Friends/acquaintances} \wedge \neg \mathit{Surprise}) = 1$
$\mathsf{P}^+(H = \mathit{Friends/acquaintances}|H' = \mathit{Friends/acquaintances} \wedge \mathit{Surprise}) = 0$

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

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

[7] 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.

### 3.4 Expansion: criminal case example

### 3.5 Refinement: downward

### 3.6 Refinement: cross

### 3.7 Refinement: