# Reply to Steeel and Stefansson

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

The assumption here is that $E$ is learned with certainty. After the agent learns about $E$, there is no longer any uncertainty about the truth of $E$. This assumption has been the target of extensive discussion. As is well-known, Jeffrey's conditionalization relaxes this assumption. The other assumption—and this is what we will focus on—is that $E$ and $A$ are propositions that belong to the agent's algebra of propositions. This algebra models what what 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. In this model, the agent cannot learn something they have never thought about. And what the agent learns cannot have any impact on the degree of belief about propositions that the agent 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 actively 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. 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 Bayesian provides a simple recipe to do that. It also coheres with the conservative spirit of Bayesian conditionalization. Bayesian conditionalization preserves the old probability distribution conditional on what has been learned. Reverse Bayesianism preserves, instead, the old probability distribution conditional on the old awareness state. **SHOW PICTURE BELOW**.

Reverse Bayesianism is a simple and elegant theory that manages to cope with a seemingly intractable problem for Bayesianism. Unfortunately, Steele and Stefansson (2021, Belief Revision for Growing Awareness) have provided a few compelling counterexamples. We believe their examples are ultimately successful, but they are liable to the objection that they are not genuine example of awareness growth. To block this objection, we provide simpler and more straightforward counterexamples. In addition, we believe that Steele and Stefansson's conclusion is too broad and overly pessimistic. They grant 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*. We will return to this distinction in due course. We agree only partly.

Much of the literature on awareness growth is concerned with a formal, algorithmic solution to the problem. Steele and Stefansson's argument suggests that a formal, algorithmic solution cannot be found, at least for cases of awareness refinement. We agree with this. At the same time, we think that awareness 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. The resulting formalism will also justify when Reverse Bayesianism hold and when it does not. The distinction between refinement and expansion that Steele and Stefansson draw, albeit a good first approximation, is too coarse and should be made more precise. We will see that there are cases of refinement in which Reverse Bayesianism (or a suitable variation of it) can be made to work.

## 2 Counterexamples?

The first counterexample Steele and Stefansson present 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 Styefansson, 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 that your partner and your best friend met one night secretively and without telling you (*Secretive*). There may be other propositions in the algebra, but these are the ones to focus on.

In light of the evidence *Secretive*, hypothesis *Affair* is more plausible than hypothesis *Friends/acquaintances*:

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

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.[1]

Next, the algebra changes. A new hypothesis is added which had not been 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 any of the other two hypotheses:

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

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

And, all things considered, this new hypothesis should be more likely than any of the other two:

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

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

So far so good. Reverse Bayesianism is not yet in trouble. Steele and Stefansson, 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 implies the other, if *Surprise* is more likely than *Affair* (by $*$), then *Friends/acquaintances* must

[1] If you were initially nearly certain your partner could not possisbly have an affair, even the fact they behaved very secretively or lied to your to meet one of your friends might not affect the probability of the two hypotheses. But this is besides the point.

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 Stefansson's conced, 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 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*):

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

So Awareness Rigidity is vindicated in this example. Reverse Bayesianism—at least the spirit of it, not the letter—can be salvaged.

Steele and Stefansson offer what they take to be a more definitive counterexample to Reverse Bayesianism (even in the form of 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 Styefansson, 2021, Section 5, Example 3)

Admittedly, this example is quite intricate. For analytic clarity, it can be split into two episodes of awareness growth. The first episode of awareness growth involved considering the language difficulty of the movie, as a new variable besides language and genre. Initially, the algebra contained the propositions *French* and \textit{{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*. This is a case of refinement because, first, you categorized movies by just language and genre, and then you added a further category, level of difficulty. But this refinement does not seem to bring about any change in the probabilities. There is no obvious reason why that should be so.

Next, you become aware and learn about something you did consider before, namely that the owner is simple minded. This learning triggers a change in degrees of belief that propagates to the genre of the movie via the language difficulty of the movie. The change in degrees of belief is triggered by the realization that the owner is simple-minded, which suggests a low level of language difficulty of the movie. The latter in turn suggests that the movie is more likely going to be a thriller rather than a comedy (possibly because thrillers are simpler—linguistically—than comedies).

Taken at face value, this example is a challenge to Reverse Bayesianism and Awareness Rigidity. It is not true, for example, that $\frac{\mathsf{P}(\mathit{Thriller})}{\mathsf{P}(\mathit{Comedy})} = \frac{\mathsf{P}^{+}(\mathit{Thriller})}{\mathsf{P}^{+}(\mathit{Comedy})}$ (against Reverse Bayesianism) nor is it true that $\mathsf{P}(\mathit{Thriller}) = \mathsf{P}^{+}(\mathit{Thriller}|\mathit{Thriller} \vee \mathit{Comedy})$ (against Awareness Rigidity). Since this is a case of refinement, the entire possibility space is fixed before and after awareness growth.

But this counterexample is likely to leave many unconvinced, or least confused. Since it consists of two parts—first the retirement by language difficulty and second the learning that the wonder is simple minded— one is left wondering which one of the two is essential for the counterexample? Are both necessary? Is only the second part really necessary while the first just give added context? What is going on here, exactly? Can we distill a simpler, more straightforward counterexample that only involves awareness growth without an episode of learning intertwined with it? For conceptual clarity, we should aspire to a neater and cleaner picture of awareness growth in cases of refinement.

The need for a conceptual clearer picture also applies to the first counterexample. There remains—we think—the need to further examine cases of awareness expansion. These cases consist in the addition of another proposition that was not previously in the algebra, but that was 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 in a more specific way to describe a meeting as friends. A more clear-cut case of awareness expansion would be the following. The police is investing a criminal cases. There are two suspects under investigation: Joe and Sue. They both had a motive. The evidence consists in a DNA match and information about how the crime was committed. Sue genetically matches the traces, but she 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 exmaple provided by Steele and Stefansson's the criminal case just outlined. They are both, arguably, cases of expansion, but they are also different. Does this matter for modelling awareness growth?

## 3 Bayesian Networks