

Rescuing Reverse Bayesianism: A Two-Step Model

Joe Roussos

Helsinki, 25 May 2026

Introduction: Bayesianism

This talk is about a problem for **Bayesianism**, as a theory of partial/degreed belief.

Bayesianism has two core commitments, as far as partial belief goes:

- 1 **Probabilism**: partial beliefs should be representable by a probability function over the relevant space of possibilities.
- 2 **Conditionalisation**: upon learning something, these beliefs should be updated by Bayesian conditionalisation, or a suitable generalisation such as Jeffrey conditionalisation.

Uncertainty in the Bayesian picture

P is central to the representation of uncertainty and resolutions of uncertainty (learning)

- ① Uncertainty about a possibility A : $0 < P(A) < 1$
- ② All learning involves shifts from P to Q , and requires **priors**
 - Bayes: Agents learns a proposition E , so $Q(E) = 1$ and $Q(A) = P(A|E)$
 - Jeffrey: Agents learns probability of a proposition $\pi(E) < 1$, so $Q(E) = \pi(E) < 1$ and $Q(A) = \pi(E)P(A|E) + \pi(\neg E)P(A|\neg E)$

Real life is a lot more uncertain

Often we learn things that we had no prior attitudes about whatsoever.

Moving to Sweden

When I moved, I opened the kitchen drawer at my sublet and saw this smörkniv. Before seeing it, not only did I not know Swedes used wooden butter knives, it had never occurred to me.



Real science is a lot more uncertain

Problem of New Hypotheses

Consider the original introduction of a new theory, like Einstein's General Relativity. It purports to explain old evidence, like Mercury's orbit. It makes new claims, like that gravity is due to curved spacetime.

Real science is a lot more uncertain

Problem of New Hypotheses

Consider the original introduction of a new theory, like Einstein's General Relativity. It purports to explain old evidence, like Mercury's orbit. It makes new claims, like that gravity is due to curved spacetime.

There are several problems for Bayesian confirmation theory here.

- Physicists had no priors for the theory, $P(GR)$ undefined
- Similarly, no priors for evidential relations $P(GR|Mercury)$
- Even if they existed, Bayesian theory doesn't allow for known evidence to do confirmatory work: already $P(Mercury) = 1$ before GR was proposed

Diagnosis

Our uncertainty contains both known unknowns (for which we have priors) and unknown unknowns—completely new possibilities, which we can't reasonably be said to have prior attitudes about.

This talk is about the **unknown unknowns**. Our question: **What constraints does rationality put on belief changes when we become aware of and learn about new possibilities?**

I will assume Bayesianism is a good theory of known unknowns, and consider a proposal called Reverse Bayesianism, which answers the question.

Characterizing the phenomenon

Movie (language)

Suppose you are deciding whether to see a movie at your local cinema. You know that the cinema only shows one 'international' movie each evening. Moreover, you know that the movie's language and genre will affect your viewing experience. The languages you consider are French and German and the genres you consider are thrillers and comedies. But then you realize that, owing to your poor French and German skills, your enjoyment of the movie will also depend on the level of difficulty of the language, whether high difficulty or low difficulty. [Steele and Stefánsson, 2021a, 1212] [cf. Steele and Stefánsson, 2021b, 56]

Refinement of awareness

French	FT	FC
German	GT	GC
	Thriller	Comedy

French	High	FTH	FCH
	Low	FTL	FCL
German	High	GTH	GCH
	Low	GTL	GCL
		Thriller	Comedy

Expansion of awareness

Movie (genre)

The case is as above, but instead of considering language difficulty, now suppose you realize that the genre could be drama. [Steele and Stefánsson, 2021a, 1212]

French	FT	FC
German	GT	GC
	Thriller	Comedy

French	FT	FC	FD
German	GT	GC	GD
	Thriller	Comedy	Drama

Setup

A Bayesian agent's initial state is represented by an algebra of propositions $\mathcal{F}$, and a probability function P .

$\mathcal{F}$ represents awareness, P represents beliefs. Changes to awareness will be thus be modelled as changes to the algebra.

Overall, start state $(\mathcal{F}, P)$, end state $(\mathcal{F}^+, P^+)$.

Steele and Stefánsson use a set of basic propositions $\mathbf{X}$ to generate $\mathcal{F}_{\mathbf{X}}$. $\bigvee \mathbf{X}$ denotes the "top" of algebra and represents the agent's tautologous proposition.

Philosophically, our task is to find plausible rationality constraints on P^+ .

Proposal: Reverse Bayesianism

Reverse Bayesianism [Bradley, 2017]

In the face of awareness change from $\mathcal{F}_{\mathbf{X}}$ to $\mathcal{F}_{\mathbf{X}+}$, one's beliefs P should be rigidly extended to $\mathcal{F}_{\mathbf{X}+}$.

Rigid Extension

For any P defined on $\mathcal{F}_{\mathbf{X}}$, a probability P^+ defined on $\mathcal{F}_{\mathbf{X}+}$ is called a *rigid extension* of P to $\mathcal{F}_{\mathbf{X}+}$ iff, for all $A \in \mathcal{F}_{\mathbf{X}}$,

$$P^+(A | \bigvee \mathbf{X}) = P(A)$$

Main sign: ratios of probabilities of "old" propositions are preserved across awareness changes.

Illustration: Movie (language)

In *Movie (language)*, RB implies that introducing the language difficulty partition should preserve these probabilities

$$P^+(G|C \vee T) = P(G)$$

French	FT	FC
German	GT	GC
	Thriller	Comedy

French	High	FTH	FCH
	Low	FTL	FCL
German	High	GTH	GCH
	Low	GTL	GCL
		Thriller	Comedy

Outline

1. Introduction
- 2. Steele and Stefánsson's challenge**
3. Diagnosis of the problem
4. Solution: the two-step model
5. Objections and replies

Steele and Stefánsson's challenge

One can devise examples where Reverse Bayesianism is violated without irrationality on the part of the agent in question. All we need are examples where awareness grows because an agent becomes aware of a proposition that she takes to be evidentially relevant, intuitively speaking, to the comparison of propositions of which she was already aware. For in that case, the ratio between probabilities of propositions of which the agent was already aware will not stay the same; one proposition will become more probable compared to the other, just as in ordinary cases where one learns evidence relevant to the comparison of hypotheses.[Steele and Stefánsson, 2021a, 1217-18]

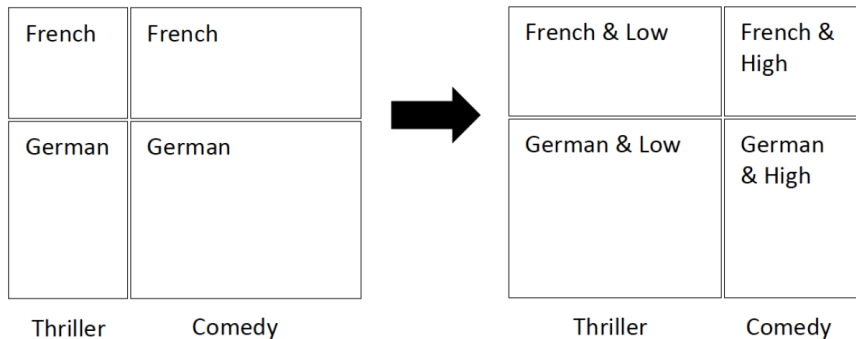
Movies, again

Recall our movie case, where you start with {French, German}, {Comedy, Thriller}. You now consider High and Low language difficulty, but in addition...

Movie (owner)

"Since it occurs to you that the owner of the cinema is quite simple-minded, you are, after this realization, much **more confident that the language of the movie will be simple** than that it will be difficult. Moreover, since you associate simple 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, 2021a, 1220]

Movies, visualised



The challenge to Reverse Bayesianism

The signature demand of Reverse Bayesianism—preserving the proportions of probabilities between known propositions—is violated here

- Start state: $\mathbf{X} = \{\text{French, German}\} \wedge \{\text{Thriller, Comedy}\}$
- End state: $\{\text{French, German}\} \wedge \{\text{Thriller, Comedy}\} \wedge \{\text{Low, High}\}$
- Reverse Bayesianism demands that

$$P^+(\text{Thriller} | \bigvee \mathbf{X}) = P(\text{Thriller})$$

But this is wrong! (say S&S) The new possibilities are *evidentially relevant* to the known possibility Thriller, so its probability must change.

Outline

1. Introduction
2. Steele and Stefánsson's challenge
- 3. Diagnosis of the problem**
4. Solution: the two-step model
5. Objections and replies

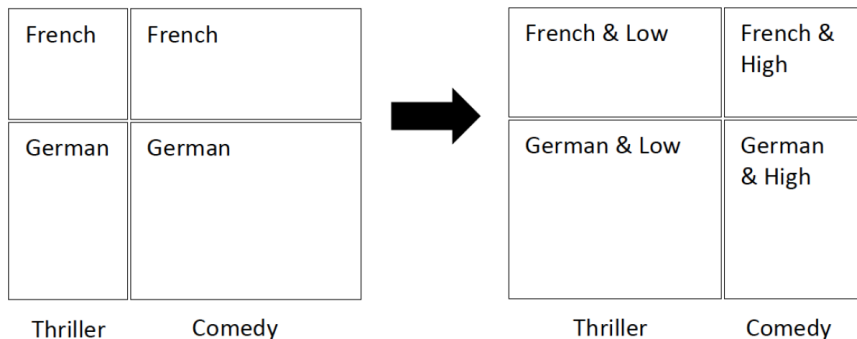
Evidential relevance

- The case turns on "evidential relevance". What is this?
- Bayesians encode this in conditional probabilities.

Evidential relevance

- The case turns on "evidential relevance". What is this?
- Bayesians encode this in conditional probabilities.
- S&S don't provide a model which includes this evidential relevance (only the diagram). Let's make it explicit!
- Case turns on a contingent fact that the agent knows: the cinema owner is simple-minded
- So the experience involved not just apprehension of the language difficulty partition, but also **grasping its association with simple-mindedness.**

Evidential relevance in *Movie (owner)*



From their figure, we see $\pi(\text{Thriller} | \text{Low})=1$ and $\pi(\text{Comedy} | \text{High})=1$. Then, the simple-mindedness link gets us $\pi(\text{Low}) > \pi(\text{High})$

Modelling the case as learning

S&S: The agent knew but neglected this fact about the cinema owner.

Modelling the case as learning

S&S: The agent knew but neglected this fact about the cinema owner.
This is OK! We are modelling it as awareness growth.

Modelling the case as learning

S&S: The agent knew but neglected this fact about the cinema owner. This is OK! We are modelling it as awareness growth. But the information about the cinema-owner has to come into the model somehow, and so I suggest that we treat it as "new" information that is "learned". This is also an as-if modelling decision.

Modelling the case as learning

S&S: The agent knew but neglected this fact about the cinema owner. This is OK! We are modelling it as awareness growth. But the information about the cinema-owner has to come into the model somehow, and so I suggest that we treat it as "new" information that is "learned". This is also an as-if modelling decision.

Benefits to this decision: seeing the connection between the *Movie* cases

- *Movie (genre)* involved just a new proposition: Drama
- *Movie (language)* involved a new partition: language difficulty
- *Movie (owner)* involves that plus evidential relevance and comparative probabilistic information

Other options?

Once we see this, we can fill in other possible combinations.

Other options?

Once we see this, we can fill in other possible combinations.

Movie (genre+)

You learn about the possibility of the cinema showing dramas by reading an announcement by the cinema: "A third of our films will now be dramas!"

So, in addition to becoming aware of the possibility Drama, you plausibly learn that $\pi(\text{Drama}) = 1/3$.

Other options?

Once we see this, we can fill in other possible combinations.

Movie (genre+)

You learn about the possibility of the cinema showing dramas by reading an announcement by the cinema: "A third of our films will now be dramas!"

So, in addition to becoming aware of the possibility Drama, you plausibly learn that $\pi(\text{Drama}) = 1/3$.

This is likely common! We often learn of a new possibility by getting information relevant to its likelihood. *Movie (owner)* is just a colourful example involving memory.

What is RB for?

Edi Karni and Marie-Louise Vierø:

*With the discovery of a contingency that he was unaware of, the decision maker's prior conception—or “model”—of the universe is falsified. When this happens, the decision maker's prior model need not be discarded; it can still provide **some guidance** for behavior in the “new” **expanded universe**. [Karni and Vierø, 2013, 2791-92]*

We're getting “some guidance” for an “expanded universe”.

What is RB for?

Karni and Vierø, again:

*According to the Bayesian paradigm, as new discoveries are made and new information becomes available, **the universe shrinks**: with the arrival of new information, events replace the prior universal state space to become the posterior state space, or universe of discourse. [...] Becoming accustomed to possibilities that were once inconceivable is part of history and our own life experience. There is a sense, therefore, in which **our universe expands** as we become aware of new opportunities. [Karni and Vierø, 2013, 2790]*

Problem: Complex experiences, like *Movie (owner)*, contain both expansions and contractions of the universe.

What is RB for?

For Richard Bradley, RB is justified by considerations of conservative belief change. He analogizes with the Rigidity condition of Jeffrey conditionalisation, noting that "conservation of the agent's relational beliefs is ensured by the rigidity of her conditional probabilities". From this, he derives the rigidity of conditional probabilities in rigid extension principle. [Bradley, 2017, 257]

Conservativeness reasoning is about **preserving what isn't affected** by a learning experience. If something is affected, then of course it shouldn't be conserved: compare Adams and Jeffrey conditioning. Proposal: **RB is for cases without new information (of any kind)**.

S&S on RB

S&S read proponents of RB as intending "a strong resemblance between awareness growth and ordinary learning and similarly between the norm for awareness growth and Bayesian conditionalisation" [Steele and Stefánsson, 2021b, 45].

S&S on RB

S&S read proponents of RB as intending "a strong resemblance between awareness growth and ordinary learning and similarly between the norm for awareness growth and Bayesian conditionalisation" [Steele and Stefánsson, 2021b, 45].

- Indeed, they dedicate a section of their book to discussing conditionalisation and its status as a norm as a warm-up to discussing RB.
- They take the relevant project to be the search for a norm that acts like conditionalisation: a one-stop shop for changes of belief, relating probabilities before and after a learning event.
- (This, no doubt, is why their paper is called "Belief revision for growing awareness".)

S&S on RB

S&S read proponents of RB as intending "a strong resemblance between awareness growth and ordinary learning and similarly between the norm for awareness growth and Bayesian conditionalisation" [Steele and Stefánsson, 2021b, 45].

- Indeed, they dedicate a section of their book to discussing conditionalisation and its status as a norm as a warm-up to discussing RB.
- They take the relevant project to be the search for a norm that acts like conditionalisation: a one-stop shop for changes of belief, relating probabilities before and after a learning event.
- (This, no doubt, is why their paper is called "Belief revision for growing awareness".)

But this is a mistake. RB is not a belief *revision* rule at all, it is a conservation condition covering only simple cases of new possibilities/"expanding the universe".

Outline

1. Introduction
2. Steele and Stefánsson's challenge
3. Diagnosis of the problem
- 4. Solution: the two-step model**
5. Objections and replies

Proposal: separate the awareness and learning stages

So what do we do with complex cases like *Movie (owner)* and *Movie (genre+)*, involving "expansions" and "contractions"?

Proposal: model in two steps.

- ① **Awareness change:** Reverse Bayesianism governs this, telling you to rigidly extend P to P^+
- ② **Belief revision:** A generalised Bayesian learning process incorporates any new information, updating P^+ to some appropriate Q

Proposal: separate the awareness and learning stages

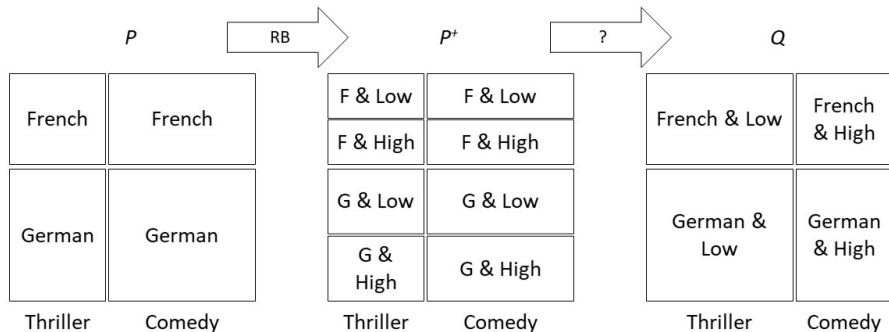
So what do we do with complex cases like *Movie (owner)* and *Movie (genre+)*, involving "expansions" and "contractions"?

Proposal: model in two steps.

- ① **Awareness change:** Reverse Bayesianism governs this, telling you to rigidly extend P to P^+
- ② **Belief revision:** A generalised Bayesian learning process incorporates any new information, updating P^+ to some appropriate Q

Technically, it has to be this way: P can't learn about propositions $A \notin \mathcal{F}_{\mathbf{x}}$! But P^+ can, because it is defined on $\mathcal{F}^+_{\mathbf{x}} \ni A$.

The proposal, visualised



What is the second step?

Any generalised belief revision procedure will do!

- **Input:** The extended prior P^+ and a "stimulus" in the form of a constraint on the posterior, e.g., $\pi(\text{Drama}) = 1/3$.
- **Output:** A posterior probability which obeys the constraint and is related to the extended prior in the right way.

Options: generalized rule-based updating (e.g., Jeffrey) [Dietrich et al., 2016] or distance-based updating [Diaconis and Zabell, 1982, Hartmann, 2025].

Common core: principle of conservative change: "It is rational to change one's degrees of belief only minimally once one learns new information".

Outline

1. Introduction
2. Steele and Stefánsson's challenge
3. Diagnosis of the problem
4. Solution: the two-step model
- 5. Objections and replies**

Steele and Stefánsson do consider this

*It might be argued that our examples are not illustrative of a simple learning event (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 (the new propositions are in the first instance evidentially irrelevant to the comparison of the old basic propositions). . . **The reason we reject** this way of conceiving of . . . our examples is that the **two-part structure is ultimately unmotivated**. The **second learning stage would be rather ad hoc, since engineered to conform to (a generalisation of) Bayesian conditionalisation**. Hence this would be seem to us to be an artificial and ad hoc way to save Reverse Bayesianism. [Steele and Stefánsson, 2021b, 57]*

Three direct responses

- 1 The examples contain new information. We want to incorporate this explicitly. You can build a one-step model if you wish, but it seems doomed: P can't learn about $A \notin \mathcal{F}$.

Three direct responses

- 1 The examples contain new information. We want to incorporate this explicitly. You can build a one-step model if you wish, but it seems doomed: P can't learn about $A \notin \mathcal{F}$.
- 2 It is S&S who wanted RB to be a one-stop shop, like conditionalisation—*not* Karni & Vierø or Bradley. If you relax the assumption that it is a total description of all changes between t_0 and t_1 then I see no reason to lump together the awareness and learning.

Three direct responses

- 1 The examples contain new information. We want to incorporate this explicitly. You can build a one-step model if you wish, but it seems doomed: P can't learn about $A \notin \mathcal{F}$.
- 2 It is S&S who wanted RB to be a one-stop shop, like conditionalisation—*not* Karni & Vierø or Bradley. If you relax the assumption that it is a total description of all changes between t_0 and t_1 then I see no reason to lump together the awareness and learning.
- 3 The final step isn't "engineered" to be Bayesian, it just is standard updating. This is the opposite of ad hoc!

Theoretical virtues of the two-step model

- ① **Coherence with prior theory:** Step 2 just is (generalised) learning. No special mechanics are required.

We have a theory of awareness change that works for cases without "evidential relevance"—Reverse Bayesianism. We have a theory of belief revision that works without awareness growth—generalised Bayesian learning. My account adds no new components, it simply shows how to combine them.

Theoretical virtues of the two-step model

- ② **Scope and unification:** It treats all these cases in a unified manner, by identifying the salient features of the case and providing a prescription for their treatment.
- *Movie (genre)*: Awareness growth, no information. Requires step 1, step 2 vacuous.
 - *Movie (genre+)*: Awareness growth, information in form of a Jeffrey-style input. Requires step 1 + 2.
 - *Movie (owner)*: Awareness growth, information in form of an inequality. Requires step 1 + 2.

Theoretical virtues of the two-step model

③ **Accuracy:** It gets the right answer!

- *Movie (genre)*: Step 1 does all the work, step 2 vacuous. We "observe RB".
- *Movie (genre+)*: Both steps active. RB's ratio "overwritten" by (Jeffrey) update.
- *Movie (owner)*: Both steps activate. More complex learning (divergence?). RB "overwritten".

S&S provide no competing account: "We take our findings to suggest that there is no general, conservative norm of belief revision for awareness growth" [Steele and Stefánsson, 2021b, 68].

How are the steps distinguished?

- They are conceptually distinct processes. Ordinary learning involves only the stimulus for step 2. Some awareness cases involve only the stimulus for step 1. S&S's cases involve both.
- In the paper I say: "the factorization is conceptual rather than psychological or temporal. The experience that brings about these steps might be a single, unitary experience, e.g., seeing the cinema's announcement about drama films."
- A reviewer objects: temporal division would make sense but is ad hoc, conceptual doesn't make sense

Temporal factorization

S&S think of RB as a process like conditionalisation, i.e, a total description of what happens between t_0 and t_1 . RB_{temp} is thus violated "if there is a time pair (t_n, t_{n+1}) where there's an awareness-growth learning event at t_n and the agent's t_{n+1} credences are not a rigid extension of their t_n credences."

Temporal factorization

S&S think of RB as a process like conditionalisation, i.e, a total description of what happens between t_0 and t_1 . RB_{temp} is thus violated "if there is a time pair (t_n, t_{n+1}) where there's an awareness-growth learning event at t_n and the agent's t_{n+1} credences are not a rigid extension of their t_n credences."

Temporal factorization would "posit that rational agents in these scenarios go through an additional temporally-distinct stage: that there's really a (t_n, t_{n+1}, t_{n+2}) where

- an awareness growth learning event occurs at t_n and t_{n+1} is a rigid extension of their t_n state,
- at which point an ordinary learning event takes place, leading to the credal state at t_{n+2} through some kind of conditionalization.

Positing a distinct learning event like this, S&S say, is odd and ad hoc."
(Reviewer)

Conceptual factorization

Reviewer, cont.: "AU wants to allow agents can rationally violate RB_{temp} ." (I think this is correct.)

Conceptual factorization

Reviewer, cont.: "AU wants to allow agents can rationally violate RB_{temp} ." (I think this is correct.)

"Can we make sense of a norm governing 'transitions' among merely conceptually distinct stages? I am not so sure."

"Whether or not we think of the change as involving an RB-respecting additional stage seems to be entirely a modelling choice, not something reflecting anything about the agent. If this is right, it is unclear what force or interest the norm is supposed to have or what would justify it."

Assessment

I think there are two complaints here.

- 1 My dismissal of S&S's ad hocness response didn't land since *they* targetted RB_{temp} .

Assessment

I think there are two complaints here.

- 1 My dismissal of S&S's ad hocness response didn't land since *they* targetted RB_{temp} .

Reply: That might be true, but if so this was a mistake on their part, brought about because Karni, Vierø and Bradley only considered cases without learning, where RB was all that happened between t_0 and t_1 . So it looked like this was a one-stop shop, although a careful reading shows otherwise.

Assessment

- ② RB_{con} looks like it is trivial, since it can't be violated and "only" occurs as a conceptual precursor to the learning stage.

Assessment

- ② RB_{con} looks like it is trivial, since it can't be violated and "only" occurs as a conceptual precursor to the learning stage.

Reply:

- The fact that RB_{con} is "overwritten" by a later learning event doesn't mean that there's no way to violate it.
- In cases without learning, an agent could straightforwardly violate it by having P^+ that is not a rigid extension of P .
- It can also be violated in learning cases. It is possible to learn something which leaves some part of the credal state unchanged. Generalized learning is very permissive, so e.g., an agent can learn a conditional probability. RB_{con} would still be "visible" in the untouched part of the credal state.

Conclusion

- ① S&S's "evidential relevance" cases snuck in new information, which RB was not intended to handle.
- ② A two-step model, with standard generalized learning as step 2, can handle these new cases.
- ③ The two-step model is coherent with prior theory, unifies the phenomena, and is accurate.
- ④ the factorization into two steps is conceptual rather than temporal, but this doesn't trivialize it.
- ⑤ The alternative is S&S's pessimism.
- ⑥ You should prefer RB instead!

References

- Richard Bradley. *Decision Theory with a Human Face*. Cambridge University Press, 2017. ISBN 978-1-107-00321-7.
- Persi Diaconis and Sandy L. Zabell. Updating Subjective Probability. *Journal of the American Statistical Association*, 77(380):822–30, 1982.
- Franz Dietrich, Christian List, and Richard Bradley. Belief revision generalized: A joint characterization of Bayes' and Jeffrey's rules. *Journal of Economic Theory*, 162: 352–371, 2016. doi: 10.1016/j.jet.2015.11.006.
- Stephan Hartmann. The Bayesian Research Programme in the Methodology of Science, or Lakatos Meets Bayes. In Roman Frigg, J. McKenzie Alexander, Laurenz Hudetz, Miklos Rédei, Lewis Ross, and John Worrall, editors, *Proofs and Research Programmes: Lakatos at 100*, pages 127–144. Springer Nature Switzerland, Cham, 2025. doi: 10.1007/978-3-031-88213-5_8.
- Edi Karni and Marie-Louise Vierø. “Reverse Bayesianism”: A Choice-Based Theory of Growing Awareness. *American Economic Review*, 103(7):2790–2810, December 2013. ISSN 0002-8282. doi: 10.1257/aer.103.7.2790.
- Katie Steele and H. Orri Stefánsson. Belief Revision for Growing Awareness. *Mind*, 130 (4):1207–1232, 2021a. doi: <https://doi.org/10.1093/mind/fzaa056>.
- Katie Steele and H. Orri Stefánsson. *Beyond Uncertainty*. Cambridge University Press, Cambridge, 2021b.