An Improved Argument for Superconditionalization

Julia Staffel

Glauber De Bona

(draft of May 2023, please cite published version in *Erkenntnis*)

Abstract

Standard arguments for Bayesian conditionalizing rely on assumptions that many epistemologists have criticized as being too strong: (i) that conditionalizers must be logically infallible, which rules out the possibility of rational logical learning, and (ii) that what is learned with certainty must be true (factivity). In this paper, we give a new factivity-free argument for the superconditionalization norm in a personal possibility framework that allows agents to learn empirical and logical falsehoods. We then discuss how the resulting framework should be interpreted. Does it still model norms of rationality, or something else, or nothing useful at all? We discuss five ways of interpreting our results, three that embrace them and two that reject them. We find one of each kind wanting, and leave readers to choose among the remaining three.

1 Introduction

Standard arguments for Bayesian conditionalization rely on assumptions that many epistemologists have criticized as being too strong, in particular: (i) that conditionalizers must be logically infallible, which precludes the possibility of rational logical learning, and (ii) that what is learned with certainty must be true (factivity), which disregards the possibility of rationally updating on a falsehood. For each of these assumptions, it has been shown that we can drop it by modifying the arguments for conditionalization, resulting in a more general version of the rule.

Pettigrew (2021a) and Rescorla (2020) have shown that we can remove the factivity assumption, resulting in arguments that show that agents who become certain of empirical falsehoods should conditionalize on them. These arguments' novelty is showing that non-conditionalizers are exposed to both accuracy dominance and Dutch book sure loss at all possible worlds, not only at worlds compatible with the learned evidence.

In separate work, Pettigrew (2021b) has demonstrated that by using a framework of personal possibilities instead of logical possibilities, the requirement of logical infallibility is removed, and agents can be modeled as updating by superconditionalization when they learn logical facts. Even though factivity is not explicitly assumed, his framework models learning via discarding the set of personally possible worlds that are incompatible with the evidence. His arguments for superconditionalization ignore accuracy and betting outcomes in worlds that are inconsistent with what the agent is certain of. We call this *pseudo-factivity*. The resulting arguments show that an agent who doesn't (super)conditionalize is *internally* irrational, since they fully believe they are in a world where they are exposed both to a Dutch book and to accuracy domination. Nevertheless, this pseudo-factive argument does not demonstrate guaranteed sure loss and accuracy domination in all possible worlds: if the agent learns something false, the actual world is not included in the pseudo-factive argument, and hence it doesn't show what happens there.

Our paper has two aims: The first is to strengthen the argument for (super)conditionalization in a personal possibility setting that allows agents to learn falsehoods. Unlike Pettigrew's argument, which assumes pseudo-factivity, our argument considers accuracy and sure loss at all worlds, including those incompatible with what the agent considers certain. This adds robustness to the agent's decision to (super)conditionalize – we show that it is the only rational update even if certainty has been misplaced. Our second aim is to explore how the resulting framework is best interpreted. Does it still model norms of rationality, or something else, or nothing useful at all? For example, it demands that agents "conditionalize" when they become certain of logical falsehoods. This might seem palatable in cases of highly complex logical reasoning, in which even a skilled reasoner could easily make an error. Yet, this version of conditionalization also requires updating when agents become certain of obvious logical falsehoods. This raises the question of whether we've gone too far - a Bayesian framework that allows agents to "learn" logical falsehoods and update on them might at best seem too soft, and at worst seriously misguided. [1] We discuss five ways of interpreting our results, three that embrace them and two that reject them. We find one of each kind wanting, and leave readers to choose among the remaining three.

2 Factivity-Free Arguments for Conditioning

Informal glosses of the conditionalization norm tend to go roughly like this: *If an agent becomes certain of some piece of evidence E, then they should update (or plan to update) their credences by making their new unconditional credences equal to their old credences that were conditional on E.* This formulation mentions the agent's attitude towards E, but omits an important detail that is usually assumed in arguments for conditionalization: that E must also be true. Rescorla (2020) has recently drawn attention to this assumption, and argues that it is desirable to provide proofs of the theorems that underlie the arguments for conditionalization (in particular, the Dutch book theorems) that don't rely on it. Pettigrew (2021a) concurs and proves a more general version of Rescorla's factivity-free Dutch book theorem, as well as a factivity-free version of the accuracy-dominance argument for conditionalization (see Briggs and Pettigrew (2020)).

The philosophical motivation for removing the factivity assumption is easy to see: conditionalization is a norm that tells rational agents how to update their credences. Rationality is commonly understood as an internalist notion, hence, agents can have high credences or beliefs in false propositions without committing a rational error (Comesana 2020). For example, a brain in a vat, or someone who is deceived by an evil demon, might have evidence that seems impeccable from their perspective, plausibly making it rational for them to conditionalize on it. The falsity of their evidence is not attributable to a failure to be a rational learner, but to having the bad fortune of being placed in an unreliable learning environment. We can think of more realistic cases as well in which learners come to acquire false information through no fault of their own. Rescorla points to instances of scientific reasoning that involve rational updates on incorrect data, among other examples.[2]

Rescorla (2020) gives a non-factive argument for conditionalization that is based on an improved version of the standard Dutch book theorem for conditionalization. His setting allows an agent to become certain of a proposition E that is not true, thus abandoning factivity. The impact of this adaptation can be better understood if we first look at an application of the standard Dutch book theorem for conditionalization.

Example 1. *Suppose the agent has the following initially coherent credences at*

t_1: $c(E) = 0.8$, $c(X\&E) = 0.4$. The agent is considering two updating rules:

- U1: $c_E(X) = 0.4$ (Don't Conditionalize)
- U2: $c_E(X) = 0.5$ (Conditionalize)

Since U1 violates conditionalization, there is a factive Dutch Book against it:

- A: at t_1, the agent buys a bet for 0.40 that returns 1 iff $X\&E$ is true;
- B: at t_1, the agent buys a bet for 0.08 that returns 0.40 iff E is false;
- C: at t_2, if the agent becomes certain of E, they sell a bet for 0.40 that returns 1 iff X is true.

Suppose that bets A, B and C take place, since the agent becomes certain of E between t_1 and t_2. The agent spent 0.48 on bets A and B at t_1, and received 0.40 back at t_2 by selling C, with a current net loss of 0.08. The agent is certain that B will not pay back and that A and C cancel each other; for the agent, X is true iff $X\&E$ is true. Thus, the agent is certain that they are losing 0.08 for sure, which is indeed the case if the evidence is true (factivity).

But what happens if E is false, unbeknownst to the agent and the bookie? The agent has again spent 0.48 on bets A and B. Since E is false, A returns nothing, and B returns 0.40, leaving the agent with a net loss of 0.08 from those two bets. Since the agent and the bookie become certain of E, despite its falsity, bet C is also placed, being sold by the agent for 0.40. The outcome then depends on whether X is true or false. If X is true, the agent must pay out 1 on bet C, leading to an overall net loss of $0.08 + 0.60 = 0.68$. But if X ends up being false, the agent keeps the selling price from bet C, leading to an overall net gain of $0.40 - 0.08 = 0.32$. Hence, failing to conditionalize does not imply that the agent loses money via a factive Dutch book.[3] Thus, factivity cannot *be simply discarded from an argument for Conditionalization, while keeping the same (factive) Dutch book theorem, or there might be other permissible updates.*

A quick and dirty way to patch up the standard argument would be to replace factivity by what we call *pseudo-factivity*: We narrow the possibilities, after becoming certain of some (true or false) evidence E, to the set of possible worlds consistent with E, say W_E. As only worlds $w \in W_E$ after updating are considered, the standard Dutch-book and accuracy-dominance theorems of the classical, factive arguments for conditionalization would be applicable.

Consider the situation in Example 1. Assuming pseudo-facticity, after becoming certain of E, the agent and the bookie rule out every world where E is false. Being aware of the factive Dutch book above, the agent knows U1 (but not U2) makes them vulnerable to sure loss in every world they are still considering as possible at t_2. Being certain of E, the agent simply ignores the possibility of E being false, where bets A, B and C could actually give them profit. Therefore, the agent, from their point of view, is compelled to adopt U2 (conditionalize).

The resulting arguments show that a non-conditionalizer should view their update as irrational, for, from their point of view, they are accuracy dominated and exposed to Dutch books. Nonetheless, these arguments would not show the pragmatic or epistemic problems of not conditionalizing from an impartial point of view, which includes worlds $w \notin W_E$ where E is false. This is undesirable, since it makes the arguments rather weak. Take one of the real-life examples that motivates Rescorla: some scientists receive data E that is, unbeknownst to them, false. How should the scientists update, given that they have become certain of E? It seems plausible that their best option is to conditionalize on E. Rescorla concurs, arguing that "even if the scientist should not have become certain of E, we can still assess how well she reallocates her other credences in light of her faulty certainty." If we hold the agent's certainty in E fixed, it's not only true from the agent's internal perspective that they should conditionalize, rather, a rational evaluation from a third-personal perspective intuitively agrees. But this third-personal perspective is left out if we assume pseudo-factivity, and just consider worlds not ruled out by the agent in the argument.

However, as an anonymous reviewer points, out, it's not obvious that on an internalist view of rationality, it matters whether there is support for conditioning from this impartial perspective in addition to the agent's own point of view. We maintain, however, that even from the agent's perspective, conditionalization stands on a stronger footing if it can be supported by an argument that doesn't assume pseudo-factivity, but considers all possible worlds. Here's why: as is widely acknowledged in discussions of preface-paradoxical cases, rational agents realize that they sometimes make mistakes, even if they are unable to spot them. Similarly, an agent who always conditionalizes on the claims they become certain of realizes that they occasionally update on false things, unbeknownst to them. They might thus wonder if always conditioning is the most desirable strategy for them to pursue in light of this. An argument that assumes pseudo-factivity only tells them that they will *think* (with credence 1) that conditioning is best in each case. By contrast, an argument that shows that conditioning is the best strategy in all possible worlds, even in those that the agent has ruled out, shows them that conditioning is in fact the most desirable updating strategy for them to implement (assuming what they become certain of is held fixed). Hence, even from the perspective of the agent, an argument for conditionalization that doesn't rely on pseudo-factivity is stronger than one that does.

While Rescorla doesn't explicitly consider a pseudo-factive modification of the standard Dutch strategy argument, his own solution cleverly avoids it, and is thus stronger. In his version of the non-factive Dutch strategy argument, when $c_E(X) \neq c(X|E)$, Rescorla suggests that the bookie make the bet at t_2 conditional on E, so that its fair relative price to the agent, who becomes certain of E^*, would be $c_{E^*}(X|E)$. The agent sees that bet as fair since either $E = E^*$, and $c_{E^*}(X|E) = c_E(X)$, or $c_{E^*}(E) = 0$, when the agent is certain that the bet will be called off. When E is not the case and the bet at t_2 is called off, the situation is analogous to the standard Dutch book for conditionalization when E is false, no bet takes place at t_2 and a suitable bet at t_1 on E guarantees the loss to the agent. When E is the case, the bet at t_2 is not called off and the difference between $c_E(X)$ and $c(X|E)$ ensures the agent's net loss, no matter whether or not $E = E^*$. In Example 1, for instance, had the bookie made the bet C on X, at t_2, conditional on E, it would have been called off in case E is false, and the agent would still have lost $0.08 = 0.48 - 0.40$ for sure, as bets A and B cost together $0.48 = 0.40 + 0.08$ and bet B would have paid 0.40 back. Rescorla's converse non-factive Dutch book theorem for conditionalization is a direct consequence of the standard version, for a sure loss in every outcome in Rescorla's scenario implies a sure loss in every scenario where the agent learns a true E.

Pettigrew (2021a) formulates different non-factive arguments for conditionalization, but he also avoids the problematic assumption of pseudo-factivity. He argues for the General Reflection Principle (GRP), a stronger norm than conditionalization, employing both Dutch book and accuracy considerations. GRP is a generalization of Van Frassen's Reflection Principle and in its weaker form demands that the current credence function, at time t_1, be a convex combination of the possible future credence functions at time t_2. Formally, the principle reads:

Weak General Reflection Principle (wGRP)(Pettigrew 2021a) Suppose c is the agent's credence function at t_1 and $c' = \langle c'_1, \ldots, c'_n \rangle$ is a tuple of credence functions they might have at t_2. Then rationality requires that there is, for each c'_i in c', a weight λ_i such that $\sum_{i=1}^{n} \lambda_i = 1$ and

$$c(-) = \sum_{i=1}^{n} \lambda_i c'_i(-)$$

Pettigrew shows how conditionalization can be directly derived from wGRP without assuming factivity. Suppose the agent will become certain of exactly one member of a partition $\{E_1, \ldots, E_n\}$ and has a planned (coherent) credence function c'_i to update to when becoming certain of each E_i, such that $c_i(E_i) = 1$. Now, if the current credences c together with $c' = \langle c'_1, \ldots, c'_n \rangle$ satisfy wGRP, then $c(X \& E_j) = c(E_j) c'_j(X)$ for any X.

The first of Pettigrew's arguments for wGRP employs Dutch strategies formed by a set of acts, which are a general form of bet. Formally, an act $A : W \to \mathbb{R}$ is a function associating a utility $A(w)$ with each possible world $w \in W$. From a probabilistic credence function c whose domain contains a proposition w representing each possible world[4], one can compute the expected utility of an act A, defined as $\sum_w c(w)A(w)$. A credence function c is said to prefer one act out of a pair if it has the higher expected utility. A pair $\langle c, c' \rangle$, formed by the prior c and a tuple c' of possible posteriors, is said to be vulnerable to a Strong Dutch strategy if there are acts A, B, A', B' such that: c prefers A to B, each c'_i in c' prefers A' to B' and $B(w) + B'(w) > A(w) + A'(w)$ for every possible word w. Now a theorem uses Dutch strategies to characterizes those $\langle c, c' \rangle$ satisfying wGRP:

Theorem 1 (Pettigrew (2021a))**.** *Let c be a probabilistic credence function and $c' = \langle c'_1, \ldots, c'_n \rangle$ be the possible future probabilistic credence functions defined* *over a set of credal objects $\mathcal{F}$ where each possible world w is represented.*

(a) If $\langle c, c' \rangle$ violates wGRP, then it is vulnerable to a Strong Dutch Strategy.

(b) If $\langle c, c' \rangle$ satisfies wGRP, then it is not vulnerable to a Strong Dutch Strategy.

In the accuracy-based argument for wGRP, Pettigrew adopts an additive, continuous, strictly proper inaccuracy measure $\mathfrak{I}$ which assigns values to credences at each possible world. This means there is a continuous strictly proper[5] scoring rule $s : [0, 1] \times [0, 1] \to [0, \infty]$ such that $\mathfrak{I}(c, w) = \sum_X s(v_w(X), c(X))$. A pair $\langle c, c' \rangle$ is then said to be accuracy dominated if there is an alternative pair, $\langle c^*, c'^* \rangle$ such that $\mathfrak{I}(c^*, w) + \mathfrak{I}(c_i'^*, w) < \mathfrak{I}(c, w) + \mathfrak{I}(c_i', w)$, for all possible worlds w and any $1 \leq i \leq n$. A theorem then states that any pair $\langle c, c' \rangle$, formed by the current credence function c and the set of possible posteriors $c' = \langle c_1', \ldots, c_n' \rangle$, is accuracy dominated if wGRP is violated, and satisfying wGRP avoids such domination:

Theorem 2 (Pettigrew (2021a))**.** *Let c be a probabilistic credence function and $c' = \langle c_1', \ldots, c_n' \rangle$ be the possible future probabilistic credence functions defined over a set of credal objects $\mathcal{F}$ where each possible world w is represented.*

(a) If $\langle c, c' \rangle$ violates wGRP, then it is accuracy dominated.

(b) If $\langle c, c' \rangle$ satisfies wGRP, then it is not accuracy dominated.

Pettigrew thus shows us two more routes towards arguing for non-factive conditionalization, both of which show that conditioning is the only rational update rule in all possible worlds, regardless of whether the agent learns a truth or a falsehood.

In the next section, we will turn to another one of Pettigrew's arguments for conditioning, which he has offered within a personal possibility framework. We will argue that it is inferior to the arguments just discussed, because it relies on pseudo-factivity.

3 Arguments for Conditioning in a Personal Possibility Framework

Dropping factivity is not the only modification to arguments for conditionalization that people have made in order to better model realistic learning scenarios. In another, unrelated strand of the literature, it has been debated how logical learning can be modeled in a Bayesian framework. Standard Bayesian models that are based on classical probabilities assume that rational agents are logically infallible. While this doesn't mean that an agent needs to know every possible logical truth, it still requires that, insofar they have any attitude at all towards a proposition, they assign credence 1 to it if it is a logical truth, and credence 0 if it is a logical falsehood. Being uncertain about, i.e., assigning middling credences to, logical truths and falsehoods is not permitted by standard Bayesian models. Also, a rational agent's credences have to correctly reflect other logical relations between the contents of their attitudes, for example, if they have credences towards two propositions X and Y, and the former entails the latter, then their credence in X can't be higher than their credence in Y. This precludes Bayesian models from representing learning experiences in which agents come to be aware of logical facts and relations that they were previously ignorant of. Yet, this kind of logical (and also mathematical) learning is common for human reasoners. Being uncertain about a logical or mathematical fact might be a failure of ideal rationality, but is not necessarily a rational defect given standards of human rationality.

A common suggestion for incorporating logical learning into a Bayesian framework is to replace logical possibilities with what is possible from the agent's perspective in formulating norms of probabilistic coherence and updating. First proposed by Hacking (1967), this idea has recently been developed further by Pettigrew (2021b). Pettigrew proposes to model an agent's growing logical awareness by replacing logical with personal possibilities as the contents of the agent's attitudes. This replacement does not preclude us from formulating Dutch book or accuracy arguments for coherence or conditionalization. The main difference is that the agent is now required to be coherent with regard to what is possible from their perspective, rather than what is logically possible. Further, Pettigrew argues that we should argue for a slightly modified version of conditionalization called "superconditionalization." Superconditionalization is slightly more general than conditionalization in the following sense: standard conditionalization assumes that there is always a proposition that the agent learns with certainty and assigns credence 1 to. Superconditionalization does not require this. Instead, an agent can directly rule out possibilities, without there being a proposition that corresponds to those possibilities, and to which the agent had assigned a credence. Pettigrew's argument for superconditioning in the personal possibility framework does not assume factivity, hence, agents can learn things that are false. Unfortunately, however, it assumes pseudo-factivity, which means that it doesn't show that conditionalizing is the only rational updating rule in all the worlds regarded as possible before the learning experience occurs. The argument only takes into account the worlds the agent considers live after E_i has been learned. We will explain how the argument works, and then motivate the need to reformulate the argument without pseudo-factivity.

Formally, Pettigrew's framework, which we mainly follow from here on, employs a set W of personally possible worlds at which each credal object from a set $\mathcal{F}$ is either true or false. Each $w \in W$ corresponds to a valuation $v_w : \mathcal{F} \to \{0, 1\}$, with $v_w(X) = 0$ if X is false at w and $v_w(X) = 1$ if X is true at w, for any $X \in \mathcal{F}$. The set of these valuations in denoted by

$W_{\mathcal{F}} = \{v_w | w \in W\}$. Note that a contradiction can be true in a given personally possible world, or a tautology false. Also, there might be worlds where X and $\neg X$ are both true, or both false, for some $X \in \mathcal{F}$. A credence function $c : \mathcal{F} \to [0, 1]$ represents the agent's numerical credences on the credal objects. In the learning scenario, the agent is about to become certain of, between t_1 and t_2, exactly one E_i from a partition $\mathcal{E} = \{E_1, \ldots, E_n\}$ of W.[6]

Note again that each E_i need not correspond to a credal object $X \in \mathcal{F}$ that is true only in worlds $w \in E_i$; that is, E_i need not be represented in $\mathcal{F}$. An updating rule c' is a function that takes each $E_i \in \mathcal{E}$ and returns a credence function c'_i, the posterior at t_2 endorsed by the rule when the agent becomes certain of E_i. Given a fixed partition $\mathcal{E} = \{E_1, \ldots, E_n\}$, we can denote an updating rule by the tuple $c' = \langle c'_1, \ldots, c'_n \rangle$, hence it can be seen as a set of possible future credences. We call a pair $\langle c, c' \rangle$, formed by a credence function at t_1 and an updating rule, a credal strategy.

Using personally possible worlds W, the (synchronic) incoherence of an agent's credence function c is defined as the existence of a set of bets it endorses that, taken together, causes loss to the agent at every world in W. A theorem by de Finetti (1974) characterizes the coherent c as those inside the convex hull of $W_{\mathcal{F}}$, denoted by $W_{\mathcal{F}}^{+}$. This motivates the following version of the Probabilism norm, parametrized by W:

Personal Probabilism(Pettigrew 2021b) Suppose c is the agent's credence function and W is the set of their personally possible worlds. Then c ought to be in $W_{\mathcal{F}}^{+}$.

A personally probabilistic c must be some weighted average of the valuations $v_w : \mathcal{F} \to [0, 1]$. That is, there must be weights $p : W \to [0, 1]$ such that $c(X) = \sum_w p(w) v_w(X)$ for all $X \in \mathcal{F}$. The function $p : W \to [0, 1]$ can be seen as a way to coherently extend c to (credal objects representing) each personally possible world, meaning that c together with p remains personally probabilistic and immune to Dutch books.

Assuming again a set W of personally possible worlds, the diachronic incoherence of a credal strategy $\langle c, c' \rangle$ is analogously defined: there is a set of bets B endorsed by c, a set B_i, for each i, endorsed by c'_i, and, for any $E_i \in \mathcal{E}$ and personally possible world $w \in E_i$, B together with B_i lead to a loss of money. To characterize the coherent credal strategies, over W, Pettigrews generalizes conditionalization to consider cases where E_i is not represented in $\mathcal{F}$:

Definition 1. $\langle c, c' \rangle$ is *superconditionalizing* if there is a function $p : W \to [0, 1]$, with $\sum_{w \in W} p(w) = 1$, such that, for all $X \in \mathcal{F}$, $c(X) = \sum_{w \in W} p(w) v_w(X)$ and for each $E_i \in \mathcal{E}$ with $\sum_{w \in E_i} p(w) > 0$:

$$c'_i(X) = \frac{\sum_{w \in E_i} p(w) v_w(X)}{\sum_{w \in E_i} p(w)}$$

Pettigrew proceeds to prove that a credal strategy $\langle c, c' \rangle$ is diachronically coherent if, and only if, $\langle c, c' \rangle$ is superconditionalizing. This yields his first argument for the following norm:

Superconditionalization(Pettigrew 2021b) Suppose c is the agent's credence function and c' is their updating rule. Then $\langle c, c' \rangle$ is superconditionalizing.

The second argument for superconditionalization is based on accuracy dominance. Assuming a continuous, additive, strictly proper inaccuracy measure $\mathfrak{I}$, Pettigrew defines that $\langle c^*, c'^* \rangle$ accuracy dominates $\langle c, c' \rangle$ if, for any $E_i \in \mathcal{E}$ and any $w \in E_i$, $\mathfrak{I}(c^*, w) + \mathfrak{I}({c'}^*_i, w) < \mathfrak{I}(c, w) + \mathfrak{I}({c'}_i, w)$. Pettigrew then proves that a credal strategy $\langle c, c' \rangle$ satisfies superconditionalization if, and only if, it is not accuracy dominated.

The similarities between the accuracy dominance argument employed in Theorem 2 and the one defined above hide a crucial difference: the worlds considered. In the argument for superconditionalization, Pettigrew assumes pseudo-factivity, evaluating sure losses and accuracy after updating only for worlds consistent with the evidence, as the agent discards the other possibilities while learning. This has a similar effect as assuming factivity, since the set of possible worlds is the learned E. His proof shows that a superconditioning credal strategy is not accuracy dominated, and this holds even if we drop factivity or consider the initial set of worlds W. Nevertheless, if we consider all initially possible worlds $w \in W$, his proof does not ensure that *only* superconditioning credal strategies will not be accuracy dominated.

Formally, the accuracy dominance mentioned in Theorem 2 holds for every pair $\langle c'_i, w \rangle$. In Pettigrew's accuracy-based argument for superconditionalization, the dominance is defined for every $\langle c'_i, w \rangle$ such that $w \in E_i$, thus ignoring, for each E_i, all worlds $w \notin E_i$. That is, the credal strategies $\langle c, c' \rangle$ and $\langle c^*, c'^* \rangle$ are compared at a world $w \in E_i$ only via c'_i and c^*_i. And in fact this detail is used in Pettigrew's proof. That is, for a non-superconditioning $\langle c, c' \rangle$, Pettigrew does not show a pair $\langle c^*, c'^* \rangle$ with $\mathfrak{I}(c^*, w) + \mathfrak{I}({c'}^*_i, w) < \mathfrak{I}(c, w) + \mathfrak{I}({c'}_i, w)$ for all i and w, including those $w \notin E_i$.

Something similar occurs in his Dutch book argument for superconditionalization. In the definition of diachronic incoherence, after the agent learns E_i, updates, and the bets B_i endorsed by c'_i take place, only net gains at worlds $w \in E_i$ are considered, due to pseudo-factivity. However, if the agent becomes certain of a false E_j, updating to c'_j, they will not necessarily engage in the bets B_i that would cause them sure loss at worlds $w \in E_i$. Again, if pseudo-factivity were dropped and we considered all possible combinations of worlds $w \in W$ and pieces of evidence $E_i \in \mathcal{E}$, superconditionalization would still avoid Dutch books, but there is no proof that only superconditioning credal strategies would do so. Indeed, a Dutch book relying on pseudo-factivity could give profit to a non-superconditionalizer at a world ignored for being incompatible with the learned evidence, as Example 1 shows.

Pettigrew's arguments are pseudo-factive: even though they do not assume the evidence E_i is true, the live possibilities while updating are reduced to the worlds consistent with E_i – as factivity would imply. Above, we argued that pseudo-factive arguments should be avoided when supporting conditionalization in cases of learning false evidence. While the pseudo-factive arguments prove that non-conditionalizers are irrational in the worlds compatible with E_i, they don't show that they are irrational in the worlds that the agent has ruled out after learning E_i. But in cases of empirical learning in a standard Bayesian framework, we wanted an argument that shows that when false evidence is learned, conditionalization is the best updating strategy not just from the perspective of the agent (and their misplaced certainty), but also from an impartial perspective that has all possible worlds in view. Both Rescorla and Pettigrew deliver such arguments in that context.

This raises the question of whether there is something special about the framework of personal possibilities that makes pseudo-factive arguments for conditionalization more appropriate. One might point out, for example, that we're only trying to model the agent's perspective, making it superfluous to attend to possibilities the agent has ruled out. We don't find this reasoning very persuasive. Even in a personal possibility framework, an agent's certainty can be misplaced. This is the case for both empirical and logical certainties. Just like in the cases discussed before, we don't just want to know whether conditioning is the only rational updating strategy from within the agent's current perspective, we also want to know if, holding the agent's certainties fixed, conditioning is the only way to go from a third-personal perspective that need not share the agent's assessment of which worlds are live possibilities. As explained above, this also reassures the agent that they should always conditionalize, even if they realize they are sometimes wrong via preface-paradox-style reasoning. If our aim is to show that conditionalization is robustly applicable even in non-ideal conditions, then it is desirable to show that it is the uniquely rational updating strategy not only in the absence of logical omniscience, but also in the presence of misplaced certainty. Avoiding pseudo-factivity is thus especially desirable in a personal possibility framework.

4 A Non-Factive Argument for Conditioning in a Personal Possibility Framework

In this section, we will show how both modifications to the standard arguments for conditioning can be combined - we can have a personal-possibility argument for superconditioning that is properly non-factive, i.e., it avoids assuming pseudo-factivity. We will show that superconditionalization can be derived from the Weak General Reflection Principle extended to personally possible worlds. The arguments for superconditionalization thus depend on those for wGRP, which we first need to adapt to the personally possible worlds framework.

First, we must reinterpret and refine the Weak General Reflection Principle in light of our framework. It suffices to assume a fixed set of credal objects $\mathcal{F}$ over which the credences are assigned. Note that wGRP does not mention a set of worlds, but only credence functions, which in principle may even violate probabilism. We can explicitly add a set of personally possible worlds W to the definition though, to refer to in the following arguments:

Weak General Reflection Principle (wGRP) Consider a set of personally possible worlds W and a set of credal objects $\mathcal{F}$, each of which is either true or false at a given world $w \in W$. Suppose $c : \mathcal{F} \to [0,1]$ is the agent's credence function at t_1 and $c' = \langle c'_1, \ldots, c'_n \rangle$ is a tuple of credence functions $c'_i : \mathcal{F} \to [0,1]$ they might have at t_2. Then rationality requires that there is, for each c'_i in c', a weight λ_i such that $\sum_{i=1}^{n} \lambda_i = 1$ and, for all $X \in \mathcal{F}$

$$c(X) = \sum_{i=1}^{n} \lambda_i c'_i(X)$$

Both arguments for wGRP by Pettigrew (2021a), employing Dutch strategy or accuracy considerations, assume probabilistic credences that are also determined for propositions representing each possible world. This brings about a problem for our framework as the set $\mathcal{F}$ of credal objects does not necessarily contain those propositions (to allow for logical learning). In the Dutch strategy argument, such assumptions are employed to determine the preference of a credence function c over a pair of acts. To address this issue, we can redefine this preference using the credence functions $p : W \to [0,1]$ that coherently extend c to all the personally possible worlds:

Definition 2. Given two acts $A : W \to \mathbb{R}$ and $B : W \to \mathbb{R}$, a personally probabilistic credence function $c : \mathcal{F} \to [0,1]$ *prefers* A to B if, for every credence function $p : W \to [0,1]$ that coherently extends c, $\sum_{w \in W} p(w)A(w) > \sum_{w \in W} p(w)B(w)$.

The Strong Dutch Strategy definition can then be applied to credence functions defined over an arbitrary set $\mathcal{F}$ of credal objects. The theorem that characterizes the pairs $\langle c, c' \rangle$ vulnerable to a Strong Dutch Strategy as those violating wGRP can now be reworked to consider an arbitrary $\mathcal{F}$.[7]

Theorem 3. *Let $c : \mathcal{F} \to [0,1]$ be a personally probabilistic credence function and $c' = \langle c'_1, \ldots, c'_n \rangle$ be the set of possible future personally probabilistic credence* *functions defined over $\mathcal{F}$.*

(a) If $\langle c, c' \rangle$ violates wGRP, then it is vulnerable to a Strong Dutch Strategy.

(b) If $\langle c, c' \rangle$ satisfies wGRP, then it is not vulnerable to a Strong Dutch Strategy.

The accuracy-based argument for wGRP put forward by Pettigrew relies on Theorem 2, which also requires some adaptation to the personally possible worlds framework, as the arbitrary set $\mathcal{F}$ of credal objects need not contain a proposition for each possible world.

Theorem 4. *Let $c : \mathcal{F} \to [0,1]$ be a personally probabilistic credence function and $c' = \langle c'_1, \ldots, c'_n \rangle$ be the set of possible future personally probabilistic credence functions defined over $\mathcal{F}$.*

(a) If $\langle c, c' \rangle$ violates wGRP, then it is accuracy dominated.

(b) If $\langle c, c' \rangle$ satisfies wGRP, then it is not accuracy dominated.

Now that we have two non-factive (and non-pseudo-factive) arguments for wGRP, considering personally possible worlds and an arbitrary set $\mathcal{F}$ of credal objects, we need to derive superconditionalization from it. The idea is that an updating rule c', with a planned c'_j for the case of becoming certain of each E_j from a given partition $\{E_1, \ldots, E_n\}$ of W, is a set of possible future (personally probabilistic) credence functions, which should satisfy wGPR in order to avoid Dutch strategies and accuracy domination. When we simply drop factivity, without replacing it by somethig else, any credal strategy $\langle c, c' \rangle$ satisfying wGRP would not be vulnerable to Dutch books or accuracy domination. For instance, the agent could plan to hold their credences fixed regardless of which evidence they become certain of, and they would still seem rational if factivity is not replaced by a suitable property[8]. But, of course, becoming certain of E_i implies some restrictions on the updated credence function, and, actually, assuming $E_1, \ldots, E_n$ are among the considered credal objects, then imposing $c'_i(E_j) = 1$ whenever $i = j$[9], given personal probabilism, suffices for wGRP to imply Conditionalization without factivity, as Pettigrew (2021a) shows[10]. However, as E_j might not be in $\mathcal{F}$ in our framework, this assumption has to be slightly modified: each c'_j must be coherently (according to personal probabilism) extendable to a credence function c^* with $c^*(E_j) = 1$. If each c'_j satisfies that property, captured by the following definition, a credal strategy $\langle c, c' \rangle$ satisfying wGRP will be superconditioning.

Definition 3. A set of credence functions $c' = \langle c'_1, \ldots, c'_n \rangle$ *respects* a partition $\{E_1, \ldots, E_n\}$ of W if, for each $1 \leq i \leq n$, there is a function $p_i : W \to [0, 1]$, with $\sum_{w \in W} p_i(w) = 1$, such that $c'_i(X) = \sum_{w \in W} p_i(w) v_w(X)$ for each $X \in \mathcal{F}$ and $\sum_{w \in E_i} p_i(w) = 1$.

Note that respecting a partition implies that all credence functions in the set c' are personally probabilistic. When each E_i is in $\mathcal{F}$, a set of credence functions $c' = \langle c'_1, \ldots, c'_n \rangle$ respects a partition $\{E_1, \ldots, E_n\}$ if, for all i and j, $c'_i(E_j) = 1$ whenever $i = j$; and personal probabilism then implies $c'_i(E_j) = 0$ for $j \neq i$. When some E_i is not in $\mathcal{F}$, respecting the partition means the agent can extend the credence functions' range to $\mathcal{F} \cup \{E_i\}$ and assign $c'_j(E_i) = 1$ for $j = i$ without violating personal probabilism. If c' is an update rule for the partition it respects, the agent plans to adopt a credence function when becoming certain of an E_i that is coherent with assigning credence 1 to E_i and credence 0 to the other $E_j \neq E_i$.

The next result derives superconditionalization for a credal strategy satisfying wGRP whose updating rule respects the partition for which it is defined:

Theorem 5. *Let c be a credence function. Let $c' = \langle c'_1, \ldots, c'_n \rangle$ be an updating rule for a partition $\{E_1, \ldots, E_n\}$ of W, respecting it. If the pair $\langle c, c' \rangle$ satisfies the Weak General Reflection Principle, then $\langle c, c' \rangle$ is superconditioning.*

When factivity does not hold, and the agent might become certain of some false E_j, adopting the credence function c'_j, according to their updating rule, Theorem 5 shows that not superconditioning on E_j implies Dutch book vulnerability and accuracy dominance. In fact, in the theorem we could have defined c' simply as a set of possible future credence functions instead of an updating rule for $\{E_1, \ldots, E_n\}$. In that case, the agent would not need to commit to adopt specifically c'_j when becoming certain of E_j. As long as those future credence functions respect a partition, wGRP requires them to be superconditioning on that partition.

Putting it all together, we have provided two arguments for a stronger version of wGRP, which holds for an arbitrary set of credal objects. Furthermore, we proved that if an agent has an updating rule $c' = \langle c'_1, \ldots, c'_n \rangle$ for a partition $\{E_1, \ldots, E_n\}$, and each c'_i is personally probabilistic implying $c'_i(E_i) = 1$[11], then wGRP entails superconditionalization.

5 Consequences and Responses

In the previous section, we generated an argument for superconditionalization that drops both the factivity assumption (without assuming pseudo-factivity) and swaps logical for personal possibilities. Our argument treats the cases of learning a truth and becoming certain of a falsehood in a parallel way. In both cases, the uniquely rational response to becoming certain of some E is to superconditionalize on it, regardless of whether the learned claim is empirical or logical, and regardless of whether we focus on all the worlds, including ones the agent no longer considers live, or on just the ones not ruled out by the agent.

In what follows, we will discuss how the resulting framework is best interpreted. While Rescorla argues in some detail for dropping factivity, and Pettigrew motivates the need to represent logical learning with personal possibilities, there has so far been no discussion of a model that combines both, even though this possibility is already implicit in Pettigrew's theory, as we explained. Our discussion is independent of our previous argument, in the sense that nothing we say in this section depends on accepting that Pettigrew's pseudo-factive arguments for superconditionalization should be replaced by ours.

While the two ways of modifying standard Bayesianism might seem individually compelling, one might worry that once we combine them, the resulting version of superconditionalization goes too far. For example, suppose an agent, call him Bob, is deliberating about installing a tree swing for his children. He is currently not sure if this can be done safely, so he needs to calculate whether the tree is strong enough to withstand the force generated by the swing. Suppose he does the calculation, which is well within his mathematical capabilities, but he makes an error. His result suggests the swing is safe, even though it's not. Still, our argument for superconditionalization recommends conditioning on the faulty result, which would then lead the agent to further conclude that building the swing is safe. Even by the standards of non-ideal norms of human rationality, our argument's verdict might seem overly permissive.

We will now discuss different possible responses to our results. To keep the discussion manageable, we will assume that readers are generally sympathetic to Bayesian theories of norms of rationality, and the standard arguments for supporting them, such as accuracy and Dutch book arguments. The question we're interested in is whether in fully dropping factivity and embracing personal possibilities, we've relaxed the standard framework too far. We will first discuss what can be said in favor of the results we've generated, and after that, discuss ways of pushing back on them. There are different ways in which one might embrace the results, which we will call (i) embrace completely, (ii) embrace and supplement, and (iii) embrace and reinterpret.

(i) Embrace Completely

One possible reaction is to think that we're getting things exactly right. On this view, the framework operates correctly in constraining the agent's credences only in light of the possibilities that are distinguished by the agent, regardless of how they map onto the logical possibilities. Further, the only relevant consideration in licensing an update is whether the agent has become certain of any of the possibilities (or ruled out any of them). What is actual and whether the update is based on logical or empirical information is irrelevant. Hence, this view essentially formalizes the idea that rational norms of coherence and reasoning should be entirely dependent on the agent's perspective, regardless of how empirically and logically (in)accurate their take on the world is. If we embrace this interpretation, the example is not taken to be worrisome: Bob is correct in thinking that he should decide whether to build the tree swing based on a calculation of the strength of the tree. And if his calculation shows him that the tree is strong enough, then, *from his perspective*, he should update his credences and decide accordingly. This is true even if his math is in fact mistaken, and the tree would break under the load. If we take seriously the idea that we're modeling what follows from the agent's *actual* point of view, then our framework *should* say that he ought to decide to build the swing.

One might further explain the motivation behind this response by pointing out that the agent's perspective can also be incorrect due to empirical factors. For example, suppose the agent knows that swings are safe to install in trees of species A but not species B, but he is unsure what species his tree belongs to. He hires a tree expert to advise him, but due to a mixup, the expert tells him species A, which is the wrong answer. Yet, having no reason to distrust the expert, the agent comes to think that his tree belongs to species A. Again, the agent would be advised by our framework to conditionalize on this information, and it would capture that, *from his perspective*, this is the sensible thing to do. On this view, which takes our framework to capture the rational way to reason given whatever input the agent has, the parallel between the two versions of the tree swing example is the correct result.

(ii) Embrace and Supplement

Even if we think that models that abandon factivity and embrace personal possibilities capture correctly how an agent should reason *given their perspective*, we might still want to be able to critically assess how they arrived at their perspective. It's one thing to think that *if* you believe you are the pope, you should infer from that that you're catholic. It's another to think that it's rational to begin with for you to think you are the pope (unless you are, which, gentle reader, is unlikely).

On this view, the personal probability model is an important ingredient in explaining what makes an agent's attitudes rational, but its verdicts are conditional in nature. If the agent's attitudes that serve as input for the model are rational, then the model tells the agent how to keep their attitudes coherent and update them. But the rational norms that govern inputs are external to the model. An account along these lines is defended by McHugh and Way (2018).

It's important to note that even standard Bayesian models that assume factivity and a classical logical possibility framework need to depend on such external norms to some extent. Suppose an agent becomes certain of an empirical truth simply by guessing correctly. If the agent then conditionalizes their credences on this truth, there is nothing in the standard Bayesian framework that would rule against the rationality of their attitude or reasoning. But we still want to say that it's irrational to become certain of something based on a pure guess, even if it the agent got lucky and guessed correctly. This verdict can only be delivered by a norm of rationality that is external to the Bayesian framework.

In our current setup, the role of these external norms has to be significantly expanded, since false logical and mathematical beliefs are no longer constituting a violation of the rules of the model. Hence, we need a set of rational norms to supplement our model that judge which of the agent's attitudes have been rationally formed and which ones have not. This gives the overall theory a much greater degree of flexibility than the standard Bayesian framework, because depending on the demandingness of one's views on rationality, verdicts about which empirical and logical judgments were rationally formed might vary considerably. For example, if our tree swing builder from before made a rather subtle error in his calculation, some theories of rational a priori belief might count his update as rationally permissible, while stricter theories might rule even subtle errors to be irrational. Similarly, depending on how sketchy the supposed tree expert appeared to be and what our standard for rational trust in testimony is, Bob's resulting high credence that his tree belongs to species A may or may not be considered rational.

But does the *embrace and supplement* strategy really alleviate the worry that the factivity-free personal probability framework is too permissive? We think one's answer to this question depends on how much of a contribution to a theory of rationality one expects from a normative formal model of rational credence. If one's expectations are fairly minimal, one might not worry about external norms doing too much of the heavy lifting. But for those who think that the formal model should be the central part of a theory of rational belief and updating, putting in so many constraints "by hand" won't be a satisfactory strategy.

(iii) Embrace and Reinterpret

Another possibility is to deny that these models are still normative.[12] Once we move to personal possibilities, and the model just captures what the agent takes to be appropriate inputs to their reasoning, these models are better interpreted as formal representations or *descriptions* of the agent's actual reasoning.

How compelling one finds this suggestion partly depends on how one interprets the normative force of these models. For example, Dogramaci (2018b) is worried that once we move to personal possibilities, the constraints of the framework become essentially impossible to violate. He says that "any case where it would initially *appear* someone is violating it [the additivity principle] will be ultimately better described, and *correctly* described, as a case where they are not violating it. Suppose I initially appear to violate [the additivity principle] by saying there's half a chance of rain tomorrow and half a chance of snow, and I think there's three quarters of a chance of rain or snow. Any such case will be better described as one where I turn out to think it might both rain and snow, and thus there are simply more doxastic possibilities (dreamt of in my philosophy) than it first appeared [...]." Pettigrew pushes back, claiming that as long as the cognitive processes by which we rule out personally possible worlds are not identical to the processes by which we assign credences, violations of personal probabilism are possible.

We think that, while Dogramaci's argument presents a serious challenge at the synchronic level, it is far less clear that the same reinterpretation strategy can be used to argue that the constraints imposed by superconditonalization are toothless. Take the agent from Dogramaci's example, whose credences have been charitably reinterpreted to take seriously the possibility that it might both snow and rain, so that $Cr(snow) = 0.5$, $Cr(rain) = 0.5$, and $Cr(snow \vee rain) = 0.75$. Suppose the agent learns that it's snowing. As a result, superconditioning provides some substantive constraints on the person's updated credences, for example that $Cr'(snow) = 1$, and that $Cr'(snow \vee rain) \geq 0.5$. What if the agent's updated credences diverge from this, so that, for instance, $Cr'(snow) = 1$ and $Cr'(snow \vee rain) = 0.25$? If we wanted to reinterpret the agent's personal probabilities to try to make them look coherent, we would have to go back to adjust our initial interpretation of the agent's credences, and we might have to engage in some serious gerrymandering of personally possible worlds to achieve this. While this is certainly a strategy for immunizing agents from violating norms of rationality, we're not sure whether it's *always* possible to reinterpret the agent's starting credences to make their updates seem rational. But even if it is often possible to do so, the idea that this could be a legitimate way of ascribing mental states to the agent is not very plausible. When an agent updates their credences in a way that appears to violate superconditioning, it's not clear why we shouldn't take this observation at face value, rather than conclude that they had some rather bizzare set of initial personally possible worlds and resulting credences that rationalize this update.

One's take on this matter will likely depend in part on one's view of how to attribute mental states to agents. We won't decide this here. But suppose that you find yourself siding with Dogramaci's argument that these models are lacking in normative force. If personal probability models can't be violated, this does not mean that these models are automatically well suited to be descriptive of the agent's reasoning. There are many lively debates in cognitive science and psychology about the exact heuristics and strategies that generate our performance on various reasoning tasks. Descriptive theories of human reasoning usually make specific predictions about how humans will think about particular problems or approach cognitive tasks. These theories are not just supposed to accommodate the data after the fact. If personal probability models are really as malleable as Dogramaci claims, then they are too malleable to make those substantive predictions. But if they can make substantive predictions about reasoning patterns, especially in diachronic cases, then Dogramaci's argument that the models have no normative force is unconvincing, since in that case, the model's prediction can be interpreted as a normative constraint on updating one's credences. Hence, either these models put substantive constraints on credences, especially in diachronic contexts, or they don't. If they do, then a normative interpretation is feasible. If they don't, then those models can't make interesting descriptive predictions about how agents will reason. Those who are unconvinced by the normative interpretation of personal probability models are left to conclude that these models live in the no man's land of pointless formal constructions that lack an interesting philosophical application.

We said above that we take our audience to be those who are generally sympathetic to Bayesian models of rational belief and updating. So those who are dissatisfied with the three strategies just discussed must think that we took our modifications of the standard Bayesian framework too far. We will thus now discuss reasons for (iv) rejecting the switch from logical to personal possibilities, and for (v) keeping factivity.

(iv) Reject Personal Possibilities

Swapping in personal for logical possibilities was supposed to modify Bayesian models to make them suitable for representing logical learning. But one might worry that Bayesian models are just the wrong tool for the job, and that even with the modification, we're trying to stretch them past their reasonable domain of application. How might one defend this position?

Take a standard Bayesian model and consider how it represents empirical learning. It takes a change in the agent's credences as input (this applies both in cases of standard and Jeffrey conditioning), and it outputs which new credence assignment the agent should adopt. In doing so, the model makes no reference to any sorts of reasoning steps the agent might undergo. Hence, it captures neither how the agent might arrive at their new credences nor whether their new credences are properly based on their previous attitudes. This makes sense - firstly, there is plausibly more than one permissible cognitive path towards arriving at the correctly updated attitudes. Secondly, a probabilistic model doesn't have sufficient structure to capture the nature of the basing relations between the agent's attitudes. As a result, various authors have interpreted the Bayesian framework as representing relations of *propositional* rationality, rather than relations of *doxastic* rationality. This means that the framework shows us what the rational attitudes are for the agent *to adopt*, in light of their evidence and prior credences, but it doesn't show us whether the agent's credences are rationally held in the doxastic sense (Smithies 2015; Wedgwood 2017; Dogramaci 2018a; Staffel 2019; Titelbaum 2019).

If we interpret the Bayesian framework in this way, then the standard appeal to logical possibilities makes a lot of sense. Assuming that logical and mathematical facts are knowable a priori, the agent is in a sense already in possession of the needed evidence that rationalizes the relevant credences in the standard framework. On this interpretation, it doesn't make sense to try to represent states of temporary logical ignorance in the framework if its real purpose is to show which attitudes are rationalized by the agent's evidence, regardless of whether the agent has worked this out already at the current moment. On this view, the use of personal probabilities to represent steps in logical reasoning is simply a confused repurposing of what the framework is supposed to model.

A common objection to this interpretation is that what an agent's evidence indicates to them should be somehow dependent on their cognitive abilities or recognitional capacities (Lord 2018; Turri 2010). Perhaps I have, in some sense, entailing evidence for or against Goldbach's conjecture, but it is still a stretch to say I have propositional justification for/against it, since it is completely beyond me to figure this out. This view might propose to relativize Bayesian norms to what is within the agent's cognitive reach to figure out. Still, this view can preserve the idea that the framework models what is propositionally rational for the agent (see Dogramaci (2018a, 2018b) for this kind of view). It endorses taking some steps towards de-idealizing standard Bayesianism, without endorsing the idea that it is suitable for modeling logical learning.

If, in light of these arguments, we resist the switch from logical to personal possibilities, then we avoid sanctioning positive credence assignments to logical and mathematical falsehoods as rational.[13] We can thus resist calling the miscalculating tree swing builder rational. This is the case even if we get rid of factivity and allow rational agents to update on *empirical* falsehoods, such as the misleading testimony about the tree species.

(v) Keep Factivity

Some philosophers, especially those who favor the knowledge-first program in epistemology, might balk at the idea that agents can rationally update on false evidence. Knowledge-firsters tend to argue that our evidence must be known, and that it is a violation of rational norms to become certain of, and conditionalize on a falsehood. Cases in which agents become certain of and update on falsehoods despite trying to "do everything right", like the case of the brain in the vat, or the example of Bob being misled by the arborist, are handled by saying that these norm violations are excused.[14]

Proponents of such a view might be tempted to think that Pettigrew's pseudo-factive argument discussed in section 3 is better than our argument. But the sense in which Pettigrew's argument assumes factivity doesn't capture what knowledge-firsters want. The pseudo-factive argument doesn't claim that it's irrational to supercondition on a falsehood, quite the opposite. This does not capture the knowledge-firster's claim that becoming certain of and conditioning on a falsehood is always irrational. Hence, the knowledge-firster doesn't gain anything from endorsing Pettigrew's argument as opposed to ours.

In fact, the knowledge-firster might even prefer our argument to Pettigrew's, for the following reason: knowledge-firsters tend to argue that in cases in which an agent doesn't have knowledge, but in which the agent has done everything in their power to be a knower, the best course of action for them is to reason as if they had knowledge. In this type of case, the agent would be excused for violating knowledge norms, since it is only due to external forces not in their control that they failed to follow the knowledge norms. In those cases, agents who become certain of a falsehood should update by superconditionalization, because that's what would be uniquely rational for someone who has knowledge. Our argument delivers this verdict, but Pettigrew's doesn't.[15]

Knowledge-firsters thus need to appeal to norms external to the model in order to impose a rational prohibition on becoming certain of, and conditionalizing on, falsehoods. But if they are interested in formalizing their idea that unlucky non-knowers are excused for their norm-violations if they update like knowers, they might very well prefer our more robust arguments for superconditionalization to Pettigrew's pseudo-factive arguments. The resulting position would ultimately turn out to be a version of the "embrace and supplement" strategy discussed above.

6 Conclusion

We have offered an argument for superconditionalization in a personal possibility framework, which shows that if an agent becomes certain of an empirical or logical claim, the uniquely rational updating strategy is superconditionalization, regardless of whether the learned claim is true or false. This means that we're greatly expanding the applicability of the superconditionalization norm. By using personal possibilities instead of logical ones, the norm applies to cases of logical learning, which it doesn't cover in standard Bayesian models. Further, since our argument avoids assuming pseudo-factivity, it more robustly supports superconditionalization as the *uniquely* rational updating rule than Pettigrew's argument.

Yet, one might have mixed feelings about such a far-reaching version of superconditionalization. As we saw, it applies even in cases like Bob's, who makes an avoidable mathematical error when calculating whether his tree can support a swing. Three possible reactions to this result stood out as most attractive in our discussion in section 5. Readers are invited to pick their favorite.

We think that the most promising way of *resisting* our argument is to say that the Bayesian framework is unsuitable for modeling logical learning. On this view, Bayesian models are best seen as modeling relations of propositional justification, which hold independently of whether the agent has recognized them through reasoning. Such a view might still embrace Rescorla's and Pettigrew's arguments that agents should update by (super-)conditionalization when they learn empirical falsehoods, but it would resist the switch to personal possibilities.

If we accept the switch to personal possibilities, there are two plausible interpretations of our results. The first one, "embrace completely", welcomes our expansion of superconditionalization, and interprets the resulting models as showing us what is rational from the agent's own perspective. It takes the agent's attitudes as a fixed input without passing judgment on them, and shows which reasoning moves seem rational from the agent's perspective. This interpretation is quite radical, as it doesn't make room for the idea that for certain irrational inputs, agents should not reason with them, but instead try to correct them.

A less radical interpretation suggests to "embrace and supplement" our argument. The idea here is that we supplement our models with separate, external norms for evaluating the attitudes that serve as inputs to the model, and then the formalism shows how the agent should update. This allows us to say that for an agent to be rational, their input attitudes must be rational, and they must be personally probabilistic and update by superconditionalization. This proposal is a way of spelling out McHugh and Way's account of good reasoning which says that a pattern of reasoning is good if it leads agents from fitting input attitudes to fitting output attitudes (McHugh and Way 2018). On this interpretation of our view, external norms make a significant contribution to determining whether an agent has rational credences. But even standard Bayesian views must rely on such external norms to some degree, which means that we would be merely expanding our reliance on them. One advantage of the "embrace and supplement" strategy is that it can formulate model-independent norms for rational input attitudes for both empirical and logical claims, whereas the standard Bayesian framework comes with fixed norms for rational attitudes towards logical claims.

Appendix

Lemma 1. *Let $\mathcal{F} = \{X_1, \ldots, X_m\}$ be a set of credal objects over a set of worlds W. Given real numbers $a_1, \ldots, a_m \in \mathbb{R}$, let $A : W \to [0,1]$ be an act defined as $A(w) = \sum_{i=1}^{m} a_i v_w(X_i)$ for all $w \in W$. If $p : W \to [0,1]$ coherently extends a personally probabilistic credence function $c : \mathcal{F} \to [0,1]$, then:*

$$\sum_{w \in W} p(w)A(w) = \sum_{i=1}^{m} c(X_i)a_i$$

Proof. Consider a $p : W \to [0,1]$ that coherently extends a personally probabilistic credence function $c : \mathcal{F} \to [0,1]$. For each $1 \leq i \leq m$, define an act $A_i : W \to [0,1]$ such that $A_i(w) = a_i v_w(X_i)$ for all $w \in W$. So we have that:

$$\begin{aligned}\sum_{w \in W} p(w)A(w) &= \sum_{w \in W} p(w) \sum_{i=1}^{m} A_i(w) \\ &= \sum_{i=1}^{m} \sum_{w \in W} p(w)A_i(w)\end{aligned}$$

Splitting the inner sum according to the truth value of X_i, for $1 \leq i \leq m$ it holds that:

$$\begin{aligned}\sum_{w \in W} p(w)A_i(w) &= \sum_{w, v_w(X_i)=1} p(w)A_i(w) + \sum_{w, v_w(X_i)=0} p(w)A_i(w) \\ &= \sum_{w, v_w(X_i)=1} p(w)a_i + \sum_{w, v_w(X_i)=0} p(w)0 \\ &= a_i \sum_{w, v_w(X_i)=1} p(w)\end{aligned}$$

As p coherently extends c, $\sum_{w, v_w(X_i)=1} p(w) = c(X_i)$. Finally, we can conclude that:

$$\sum_{w \in W} p(w)A(w) = \sum_{i=1}^{m} c(X_i)a_i$$

□

Theorem 3 . *Let $c : \mathcal{F} \to [0,1]$ be a personally probabilistic credence function and $\mathrm{c}' = \langle c'_1, \ldots, c'_n \rangle$ be the set of possible future personally probabilistic credence functions defined over $\mathcal{F}$.*

(a) If $\langle c, \mathrm{c}' \rangle$ violates wGRP, then it is vulnerable to a Strong Dutch Strategy.

(b) If $\langle c, \mathrm{c}' \rangle$ satisfies wGRP, then it is not vulnerable to a Strong Dutch Strategy.

Proof. (a) Supposing $\mathcal{F} = \{X_1, \ldots, X_m\}$, any credence function $\hat{c} : \mathcal{F} \to [0,1]$

can be viewed as a vector $\langle \hat{c}(X_1), \ldots, \hat{c}(X_m) \rangle \in \mathbb{R}^m$. If $\langle c, c' \rangle$ violates wGRP, then c is not in the convex hull of the vectors $c'_1, \ldots, c'_n$. Thus, by the Separating Hyperplane Theorem, there are numbers $a, a' \in \mathbb{R}$ and a vector $b \in \mathbb{R}^m$ such that, for each $c'_j \in c'$:

$$\sum_{i=1}^{m} c(X_i) b_i < a < a' < \sum_{i=1}^{m} c'_j(X_i) b_i$$

Now consider the constant acts $A : W \to [0,1]$ and $A' : W \to [0,1]$ such that $A(w) = a$ and $A'(w) = -a'$ for all $w \in W$. Furthermore, consider the act $B : W \to [0,1]$ defined in the following way: $B(w) = \sum\limits_{i=1}^{m} b_i v_w(X_i)$. That is, for each world $w \in W$, determine those X_i that are true and sum the corresponding b_i to obtain $B(w)$. If all $X_i \in \mathcal{F}$ are false in w, then $B(w) = 0$. Now, define the act $B' : W \to [0,1]$ via $B'(w) = -B(w)$ for all $w \in W$. By Lemma 1, for any $p : W \to [0,1]$ that coherently extends c we have that $\sum_w p(w)B(w) = \sum\limits_{i=1}^{m} c(X_i) b_i$. As $\sum_w p(w)A(w) = a$, c prefers A to B. Analogously, by Lemma 1, each c'_j prefers A' to B'. Finally, note that $B(w) + B'(w) = 0 > a - a' = A(w) + A'(w)$ for any $w \in W$, therefore $\langle c, c' \rangle$ is vulnerable to a Strong Dutch Strategy.

(b) Assume there are weights $\lambda_j \in [0,1]$, summing up to one, such that $c(X) = \sum_j \lambda_j c'_j(X)$ for all $X \in \mathcal{F}$. To prove by contradiction, suppose there are acts A, A', B, B' such that c prefers A to B, each c_j prefers A' to B', but $A(w) + A'(w) < B(w) + B'(w)$ at any $w \in W$. For each c'_j, consider a credence function $p'_j : W \to [0,1]$ that coherently extends it, thus preferring A' to B'. Defining $p : W \to [0,1]$ via $p(w) = \sum_j \lambda_j p'_j(w)$ for all $w \in W$, we have that $\langle p, p' \rangle$ satisfies wGRP. Note that p coherently extends c, hence preferring A to B. Consequently, $\langle p, p' \rangle$ is vulnerable to a Strong Dutch Strategy, which contradicts Theorem 1(b). □

Theorem 4 . *Let $c : \mathcal{F} \to [0,1]$ be a personally probabilistic credence function* *and* $\mathbf{c}' = \langle c'_1, \ldots, c'_n \rangle$ *be the set of possible future personally probabilistic credence functions defined over* $\mathcal{F}$*.*

(a) If $\langle c, \mathbf{c}' \rangle$ *violates wGRP, then it is accuracy dominated.*

(b) If $\langle c, \mathbf{c}' \rangle$ *satisfies wGRP, then it is not accuracy dominated.*

Proof. (a) See the $(\rightarrow)$-part of the proof of Theorem 2 (Pettigrew 2021a), just interpreting W as a set of personally possible worlds.

(b) Suppose $\langle c, \mathbf{c}' \rangle$ satisfies wGRP, so that there are $\lambda_1, \ldots, \lambda_n \in [0, 1]$ such that $\sum_j \lambda_j = 1$ and $c(X) = \sum_{j=1}^{n} \lambda_j c_j(X)$ for any $X \in \mathcal{F}$. To prove by contradiction, assume $\langle c, \mathbf{c}' \rangle$ is accuracy dominated: there are credence functions $c^*, c^*_1, \ldots, c^*_n$, defined on $\mathcal{F}$, such that $\mathfrak{I}(c, w) + \mathfrak{I}(c'_j, w) > \mathfrak{I}(c^*, w) + \mathfrak{I}(c^*_j, w)$ for all $w \in W$ and $1 \leq j \leq n$. Since accuracy is measured with an additive strictly proper $\mathfrak{I}$, there is a strictly proper scoring rule s such that, for any credence function $\hat{c}$, $\mathfrak{I}(\hat{c}, w) = \sum_{X \in \mathcal{F}} s(v_w(X), \hat{c}(X))$. For s is strictly proper, we have that, for any $X \in \mathcal{F}$ and any $1 \leq j \leq n$:

$$c(X)s(1, c(X)) + (1 - c(X))s(0, c(X)) \leq c(X)s(1, c^*(X)) + (1 - c(X))s(0, c^*(X)) \quad (1)$$

$$c'_j(X)s(1, c'_j(X)) + (1 - c'_j(X))s(0, c'_j(X)) \leq c'_j(X)s(1, c^*_j(X)) + (1 - c'_j(X))s(0, c^*_j(X)) \quad (2)$$

If we replace those $c(X)$ out of the scope of $s(.)$ by $\sum_j \lambda_j c_j(X)$ in Expression (1) (note also that $\sum_j \lambda_j = 1$) and, in Expression (2), multiply both sides by λ_j before summing for all j, we obtain, respectively:

$$\begin{aligned}
&\sum_j \lambda_j c'_j(X)s(1, c(X)) + \sum_j \lambda_j (1 - c'_j(X))s(0, c(X)) \leq \\
&\sum_j \lambda_j c'_j(X)s(1, c^*(X)) + \sum_j \lambda_j (1 - c'_j(X))s(0, c^*(X)) \qquad (3) \\
&\sum_j \lambda_j \left[c'_j(X)s(1, c'_j(X)) + (1 - c'_j(X))s(0, c'_j(X))\right] \leq
\end{aligned}$$

$$\sum_j \lambda_j \left[c'_j(X)s(1, c^*_j(X)) + (1 - c'_j(X))s(0, c^*_j(X))\right] \tag{4}$$

Grouping the summations in j in each side of Expression (3), it can be added to Expression (4) to obtain:

$$\begin{aligned}&\sum_j \lambda_j \left[c'_j(X)(s(1, c(X)) + s(1, c'_j(X))) + (1 - c'_j(X))(s(0, c(X)) + s(0, c'_j(X)))\right] \leq \\ &\sum_j \lambda_j \left[c'_j(X)(s(1, c^*(X)) + s(1, c^*_j(X))) + (1 - c'_j(X))(s(0, c^*(X)) + s(0, c^*_j(X)))\right]\end{aligned} \tag{5}$$

Since $\langle c, c' \rangle$ is accuracy dominated by c^* and $\langle c^*_1, \ldots, c^*_n \rangle$, $\mathfrak{I}(c, w) + \mathfrak{I}(c'_j, w) > \mathfrak{I}(c^*, w) + \mathfrak{I}(c^*_j, w)$ for all $w \in W$ and $1 \leq j \leq n$. Multiplying each side of this inequality by $\lambda_j p_j(w)$, for a p_j that coherently extends c'_j, and summing for all $w \in W$ and all $1 \leq j \leq n$, we obtain:

$$\sum_j \lambda_j \sum_w p_j(w)\left[\mathfrak{I}(c, w) + \mathfrak{I}(c'_j, w)\right] > \sum_j \lambda_j \sum_w p_j(w)\left[\mathfrak{I}(c^*, w) + \mathfrak{I}(c^*_j, w)\right] \tag{6}$$

Recall that, for any $\hat{c} : \mathcal{F} \to [0, 1]$, $\mathfrak{I}(\hat{c}, w) = \sum_{X \in \mathcal{F}} s(v_w(X), \hat{c}(X))$. Thus, for any $\hat{c}$, $\mathfrak{I}(\hat{c}, w)$ can be rewritten as:

$$\begin{aligned}\mathfrak{I}(\hat{c}, w) &= \sum_{X \in \mathcal{F}} v_w(X)s(1, \hat{c}(X)) + \sum_{X \in \mathcal{F}} (1 - v_w(X))s(0, \hat{c}(X)) \\ &= \sum_{X \in \mathcal{F}} v_w(X)s(1, \hat{c}(X)) - \sum_{X \in \mathcal{F}} v_w(X)s(0, \hat{c}(X)) + \sum_{X \in \mathcal{F}} s(0, \hat{c}(X))\end{aligned} \tag{7}$$

If $\hat{c}$ is fixed, the first two summations in Expression (7) can be viewed as acts in the format $\sum_i a_i v_w(X_i)$. Hence, applying Lemma 1 to $\sum_w p_j(w)\mathfrak{I}(c, w)$ yields:

$$\sum_{X \in \mathcal{F}} c'_j(X)s(1, c(X)) - \sum_{X \in \mathcal{F}} c'_j(X)s(0, c(X)) + \sum_w p_j(w) \sum_{X \in \mathcal{F}} s(0, c(X))$$

As $\sum_w p_j(w) = 1$, a bit of algebraic manipulation results in:

$$\sum_{X \in \mathcal{F}} \left[c'_j(X)s(1, c(X)) + (1 - c'_j(X)s(0, c(X))\right]$$

Analogously, we can apply Lemma 1 to each $\sum_w p_j(w)\mathfrak{I}(\cdot, w)$ resulting from Expression (6), obtaining:

$$\sum_j \sum_{X \in \mathcal{F}} \lambda_j \left[c'_j(X)(s(1, c(X)) + s(1, c'_j(X))) + (1 - c'_j(X))(s(0, c(X)) + s(0, c'_j(X)))\right] >$$
$$\sum_j \sum_{X \in \mathcal{F}} \lambda_j \left[c'_j(X)(s(1, c^*(X)) + s(1, c^*_j(X))) + (1 - c'_j(X))(s(0, c^*(X)) + s(0, c^*_j(X)))\right] \quad (8)$$

But note that this is just the negation of Expression (5) summed for all $X \in \mathcal{F}$, which is a contradiction, completing the proof. □

Theorem 5 . *Let c be a credence function. Let $c' = \langle c'_1, \ldots, c'_n \rangle$ be an updating rule for a partition $\{E_1, \ldots, E_n\}$ of W, respecting it. If the pair $\langle c, c' \rangle$ satisfies the Weak General Reflection Principle, then $\langle c, c' \rangle$ is superconditioning.*

Proof. As c' respects the partition $\{E_1, \ldots, E_n\}$, for each $1 \leq i \leq n$ there is a function $p_i : W \to [0, 1]$, with $\sum_{w \in W} p_i(w) = 1$, such that $c'_i(X) = \sum_{w \in W} p_i(w)v_w(X)$ for all $X \in \mathcal{F}$ and $\sum_{w \in E_i} p_i(w) = 1$. wGRP implies that, for some $\lambda_1, \ldots, \lambda_n \in [0, 1]$ with $\sum_{i=1}^{n} \lambda_i = 1$, we have that $c(X) = \sum_{i=1}^{n} \lambda_i c'_i(X)$ for all $X \in \mathcal{F}$. Thus, $c(X) = \sum_{i=1}^{n} \lambda_i \sum_{w \in W} p_i(w)v_w(X) = \sum_{w \in W} \sum_{i=1}^{n} \lambda_i p_i(w)v_w(X)$ for all $X \in \mathcal{F}$. Let the function $p : W \to [0, 1]$ be such that $p(w) = \sum_{i=1}^{n} \lambda_i p_i(w)$ for all $w \in W$. Note that $\sum_{w \in W} p(w) = 1$.

Now consider an element E_j of the partition. We have, for all $X \in \mathcal{F}$, that $\sum_{w \in E_j} p(w)v_w(X) = \sum_{w \in E_j} v_w(X) \sum_{i=1}^{n} \lambda_i p_i(w)$. Given that, for any $w \in E_j$, $p_i(w) = 0$ whenever $i \neq j$, as the partition is respected, it follows that $\sum_{w \in E_j} v_w(X) \sum_{i=1}^{n} \lambda_i p_i(w) = \sum_{w \in E_j} v_w(X)\lambda_j p_j(w) = \lambda_j \sum_{w \in W} v_w(X)p_j(w)$. For all

$w \in E_j$, we also have that $\sum\limits_{w \in E_j} p(w) = \sum\limits_{w \in E_j} \sum\limits_{i=1}^{n} \lambda_i p_i(w) = \sum\limits_{w \in E_j} \lambda_j p_j(w)$, thus $\sum\limits_{w \in E_j} p(w) = \lambda_j \sum\limits_{w \in E_j} p_j(w) = \lambda_j$. Finally, for each $1 \leq j \leq n$ with $\sum\limits_{w \in E_j} p(w) = \lambda_j > 0$, we obtain, for all $X \in \mathcal{F}$:

$$c'_j(X) = \sum_{w \in W} p_j(w) v_w(X) = \frac{\lambda_j \sum\limits_{w \in W} p_j(w) v_w(X)}{\lambda_j} = \frac{\sum\limits_{w \in E_j} p(w) v_w(X)}{\sum\limits_{w \in E_j} p(w)}$$

Consequently, $\langle c, c' \rangle$ is superconditioning. □

Notes

[1]Henceforth, we will use "learn" and "become certain of" interchangeably. Hence, when we say that an agent learns something, what they learn can be true or false.

[2]One objection is worth mentioning here, although we won't discuss it in detail, as it would lead us away from our main train of thought. It says that becoming *certain* of empirical evidence is always irrational. Proponents of the regularity principle argue that we should at most invest high credence in any empirical propositions, but not credence 1. This makes Jeffrey conditionalization the only relevant update rule, thus omitting the need for a factivity-free version of conditioning. We think that the case for regularity is unconvincing. For discussion, see Rescorla (2020) and especially Hájek (2012). Also, proponents of contextualist versions of Bayesianism give good reasons to avoid regularity as a requirement (Greco 2017; Salow 2019).

[3]A similar problem would occur if E was true but the agent became certain of a $E' \neq E$. In that case, the bet at t_2 would not occur, for the bookie would also be certain of E', but the standard Dutch book relies on it, for E is true.

[4]Abusing the notation, we use w for both a possible world and the proposition that is true only there.

[5]s is strictly proper if only $x = p$ minimizes $ps(1, x) + (1 - p)s(0, x)$ for any $p \in [0, 1]$.

[6]We assume in our argument below that W is kept fixed after learning, to avoid pseudo-factivity, while in Pettigrew's argument this set is narrowed to the learned E_i, which makes his arguments pseudo-factive.

[7]Proofs can be found in the Appendix.

[8]Pseudo-factivity would do the job, but we are trying to avoid it.

[9]Note this property is, given (Personal) Probabilism, implied by but weaker than restricting the set of worlds W to those compatible with the learned E_j.

[10]Also Rescorla (2020) assumes a similar assumption, for an agent that became certain of an E_i to accept any bet conditional on a $E_j \neq E_i$.

[11]If $E_i \notin \mathcal{F}$, c'_i implies $c'_i(E_i) = 1$ if it is the only credence that personally probabilistically extends c'_i to E_i.

[12]Thanks to Kenny Easwaran for suggesting that we discuss this option.

[13]Except perhaps in cases in which it is beyond an agent's cognitive capacities to assign the correct credences.

[14]This strategy is defended by Williamson (forthcoming). For a critical discussion, see Greco (forthcoming).

[15]Notice that moving to the standard arguments for conditioning would not help much here. In the standard framework, it's irrational for the agent to become certain of a logical falsehood. But when an agent becomes certain of an empirical falsehood, the arguments give the same verdict as Pettigrew's version from section 3: conditionalizing on the falsehood avoids Dutch-bookability and accuracy dominance, but it's not shown that it's the unique strategy with these properties.

References

Briggs, R. A. and R. Pettigrew (2020). An accuracy-dominance argument for conditionalization. *Noûs 54(1)*, 162–181.

Comesana, J. (2020). *Being Rational and Being Right.* Oxford University Press.

de Finetti, B. (1974). *Theory of probability.* John Wiley and Sons, Chichester.

Dogramaci, S. (2018a). Rational credence through reasoning. *Philosophers' Imprint 18*.

Dogramaci, S. (2018b). Solving the problem of logical omniscience. *Philosophical Issues 28*(1), 107–128.

Greco, D. (2017). Cognitive mobile homes. *Mind 126*(501), 93–121.

Greco, D. (2021). Justifications and excuses in epistemology. *Noûs 55*(3), 517–537.

Hacking, I. (1967). Slightly more realistic personal probability. *Philosophy of Science 4*, 311–325.

Hájek, A. (2012). Is strict coherence coherent? *Dialectica 66*(3), 411–424.

Lord, E. (2018). *The Importance of Being Rational*. Oxford, UK: Oxford University Press.

McHugh, C. and J. Way (2018). What is good reasoning? *Philosophy and Phenomenological Research*, 153–174.

Pettigrew, R. (2021a). Bayesian updating when what you learn might be false. *Erkenntnis*, 1–16.

Pettigrew, R. (2021b). Logical ignorance and logical learning. *Synthese 198*(10), 9991–10020.

Rescorla, M. (2020). An improved dutch book theorem for conditionalization. *Erkenntnis.*, 1–29.

Salow, B. (2019). Elusive externalism. *Mind 128*(510), 397–427.

Smithies, D. (2015). Ideal rationality and logical omniscience. *Synthese 192*(9), 2769–2793.

Staffel, J. (2019). *Unsettled Thoughts: A Theory of Degrees of Rationality*. Oxford University Press.

Titelbaum, M. (2019). Return to reason. In A. S.-P. M. Skipper (Ed.), *Higher-Order Evidence: New Essays*. Oxford: Oxford University Press.

Turri, J. (2010). On the relationship between propositional and doxastic justification. *Philosophy and Phenomenological Research 80*(2), 312–326.

Wedgwood, R. (2017). *The Value of Rationality*. Oxford University Press.

Williamson, T. (forthcoming). Justifications, excuses, and sceptical scenarios. In F. Dorsch and J. Dutant (Eds.), *The New Evil Demon*. Oxford: Oxford University Press.