UPDATING INCOHERENT CREDENCES - EXTENDING THE DUTCH STRATEGY ARGUMENT FOR CONDITIONALIZATION

Glauber De Bona

Universidade de São Paulo

glauber.bona@usp.br

Julia Staffel

University of Colorado at Boulder

staffel.julia@gmail.com

July 24, 2021

Abstract

In this paper, we ask: how should an agent who has incoherent credences update when they learn new evidence? The standard Bayesian answer for coherent agents is that they should conditionalize; however, this updating rule is not defined for incoherent starting credences. We show how one of the main arguments for conditionalization, the Dutch strategy argument, can be extended to devise a target property for updating plans that can apply to them regardless of whether the agent starts out with coherent or incoherent credences. The main idea behind this extension is that the agent should avoid updating plans that increase the possible sure loss from Dutch strategies. This happens to be equivalent to avoiding updating plans that increase incoherence according to a distance-based incoherence measure.

1   Introduction

A central question in Bayesian epistemology is how agents should update their credences when they learn new information with certainty. The most popular answer is that they

1

should conditionalize. Conditionalization has been defended using a variety of different arguments, most prominently Dutch strategy arguments (Lewis, as published by Teller (1973)) and accuracy-based arguments (Briggs & Pettigrew, 2020; Greaves & Wallace, 2006; Oddie, 1997). A major limitation of conditionalization is that it is only defined for agents whose credences are probabilistic. This means that the rule is silent about how to best update from incoherent starting points. Nonetheless, even in those cases, some updating plans seems to be more adequate than others, as in the following example.

Example 1. *Consider two agents, Jane and Clara, whose initial credences are identical and incoherent: $cr(A\&B) = 0.4$, $cr(\neg A\&B) = 0.4$, $cr(B) = 0.9$ and $cr(\neg B) = 0.1$. Upon learning B, Jane's updating plans include $cr_B(A) = 0.5$ and $cr_B(\neg A) = 0.5$, while Clara's include $cr_B(A) = 0.3$ and $cr_B(\neg A) = 0.7$. Even though both are in principle going to be coherent after learning B, there is some intuitive pull to judge Jane's updating plans to be more aligned with her initial credences than Clara's.*

Our aim in this paper is to explore whether we can devise a target property for updates that preserves the spirit of conditionalization, but that can apply both to agents with initially coherent and with initially incoherent credences.[1] We show how this can be done by utilizing the Dutch strategy argument for conditionalization. The main idea behind our proposal is that an agent should plan to update their credences in a way that doesn't increase the possible sure loss from Dutch strategies. This happens to be equivalent to avoiding updating plans that increase incoherence according to a distance-based incoherence measure. The resulting way of evaluating updates recommends updating by conditionalization to coherent agents, and updating in a similar way that we call *tolerant conditionalization (TC)* to incoherent agents.[2]

Our discussion will proceed as follows: In section 2, we present the standard Dutch strategy argument for conditionalization, and explain why it is an undesirable limitation that this update norm is only defined for initially coherent agents. In section 3, we review some existing distance-based approaches to measuring incoherence that track Dutch book loss. We explain how incoherent agents who have fixed information can augment their existing credences without becoming more incoherent. In section 4, we show how this idea can be applied to cases in which incoherent agents learn new information. Dutch strategy loss can be measured as the incoherence of the agent's initial unconditional credences, extended with their updating plans. We call the additional incoherence stemming from

---

[1] For different approaches to this issue, see Gilboa and Schmeidler (1993; 1994) and Babic (2020). Gilboa and Schmeidler propose update rules for non-additive probability measures whose application need not result in additive ones. Babic explores what happens if incoherent agents use a version of standard conditionalization without renormalizing, and argues that the results tell against norms that recommend approximating ideally rational credences. We take the opposite route and argue from these approximation norms to a different way of extending conditionalization to incoherent agents, yielding coherent updated credences.

[2] Thanks to Justin Snedegar for suggesting the name.

this extension *plan incoherence*. According to *tolerant conditionalization*, an incoherent agent should plan to update their credences upon learning some piece of new information with certainty in a way that minimizes their *plan incoherence*. In section 5, we show how our initial models can be enriched by incorporating agents' conditional credences in addition to their unconditional credences and updating plans. In section 6, we discuss options for using alternative incoherence measures, and also how our results can impact the accuracy of an agent's credences.

2   The Dutch Strategy Argument for Conditionalization and its Limitations

In Bayesian epistemology, the standard answer to the question of how an agent should update their credences upon becoming certain of a new piece of information is "Conditionalize!". More specifically, when constraining the *actual* updated credences, the norm goes like this:[3]

Conditionalization: When new evidence A is acquired with certainty (and no other evidence is acquired), the resulting credence in every claim B is equal to its previous credence conditional on A, so $cr_{new}(B) = cr_{old}(B|A)$, assuming that $cr_{old}(A) > 0$.

Conditional credences are standardly required to obey the ratio formula according to the norms of ideal Bayesian rationality:

Ratio Formula: The conditional credence of B given A, written $cr(B|A)$, equals $cr(A\&B)/cr(A)$, assuming that $cr(A) > 0$.

The norm of conditionalization has been defended with a variety of arguments. The first argument in its favor was offered by David Lewis.[4] It is a *Dutch strategy argument* for conditionalization, which shows that if an agent doesn't update by conditionalization, they are vulnerable to a sequence of bets that will lose them money no matter what. More recently a couple of accuracy-based arguments for conditionalization have been offered. Oddie (1997) as well as Greaves and Wallace (2006) have argued that the expected accuracy of conditionalization is higher than the expected accuracy of any other updating rule. Briggs and Pettigrew (2020) have argued that any non-probabilistic credal strategy (a combination of initial credences and updating plans) is accuracy-dominated by a probabilistic, conditionalizing strategy, while probabilistic, conditionalizing strategies are undominated.[5] Some philosophers prefer accuracy-based arguments for epistemic norms

---

[3]We distinguish between two versions of conditionalization, following Pettigrew (2020), where this is called "Actual Conditionalization"

[4]The argument is standardly attributed to David Lewis, but first appeared in print in a paper by Paul Teller (who credits Lewis) (Teller, 1973).

[5]Diaconis and Zabell (1982) have furthermore shown that conditionalizing a credence assignment

to Dutch book arguments, because they want to avoid appealing to pragmatic considerations in justifying epistemic norms. Yet, as, for instance, Christensen (2004) has argued, less literal interpretations of Dutch book arguments can avoid some of these objections. In principle, one can endorse both types of arguments insofar as they support the same conclusions.

Our aim in this paper is to extend the Dutch strategy argument to come up with a desirable target property for updates based on incoherent starting points. Hence, we will focus on Dutch strategies for now, but we will return to the question of how our results connect to accuracy-based justifications of conditionalization in section 5. It has recently been shown that the standard Dutch strategy argument for conditionalization must assume that the agent's evidence consists in learning that exactly one member of a partition is true. We will adopt this assumption throughout the paper.[6]

Here is a simple example of how a non-conditionalizing agent can be exploited with a Dutch strategy.

Example 2. *Suppose an agent is to learn the truth of one of the members of the partition* $\{B, \neg B\}$ *and begins with the following probabilistic starting credences:*

$cr(A\&B) = 0.4$, $cr(\neg A\&B) = 0.4$, $cr(B) = 0.8$ *and* $cr(\neg B) = 0.2$.

This commits them to the conditional credence $cr(A|B) = 0.5$, *insofar as they obey the ratio formula.*

---

defined over a partition of propositions on E is equivalent to moving to the closest coherent credence assignment that assigns $cr(E) = 1$ according to the Kullback-Leibler divergence. Unfortunately, their result only applies to partitions, not to credence assignments over arbitrary sets of propositions.

[6]Proponents of evidential externalism want to reject the partitionality assumption. They argue that in cases in which it is not transparent to the agent what their evidence is, the partitionality assumption fails, which means that a crucial premise in the Dutch strategy argument is unavailable (Das, 2020; Gallow, 2019a). The problem also applies to the expected accuracy argument for conditionalization, as Schoenfield (2017) shows, as well as the accuracy dominance arguments by Briggs and Pettigrew (2020) and Nielsen (2021). It is a matter of current debate how to respond to these results, but we can see two main strategies in the recent literature. The first strategy is to modify how we should understand the relevant notion of evidence.

(1a) For example, Schoenfield (2017) argues, in line with earlier results by Shafer (1985) and Hild (1998), that agents must update on the claim that they have learned some piece of evidence E, rather than E itself, which is a way of guaranteeing that there are no externalist counterexamples.

(1b) Alternatively, we might employ a contextualist (or relativist/expressivist) interpretation of what counts as the agent's evidence. Salow (2019), picking up a similar idea in Greco (2017), offers a view on which the referent of the term "evidence" can never fail to satisfy certain internalist principles in a given context.

Both (1a) and (1b) are compatible with the approach in this paper. The second strategy is to devise alternative, externalist-friendly arguments for either conditionalization or another updating rule.

(2a) For example, Gallow (2019b) and Zendejas Medina (2021) set out to devise such alternative updating strategies. Unfortunately, their proposals rely on expectation-based arguments, which are difficult to generalize to agents with incoherent starting credences.

(2b) Rescorla (2020) discusses how the factivity assumption in standard Dutch strategy arguments can be relaxed. His discussion might give us some hints about how to develop future proposals that will be compatible with the approach taken in this paper.

We are currently undecided on which of these strategies is most likely to be successful, but at least the first one would be compatible with our approach here.

Hence, if they were to conditionalize upon learning that B, they would have to update their credence in A to 0.5. Now, let's suppose they don't plan to conditionalize, but to adopt a 0.4 credence in A upon learning that B, which is the non-normalized posterior $cr(A\&B) = 0.4$, so $cr_B(A) = 0.4$. In that case, before knowing whether B is true, the bookie will sell the agent the following two bets:

Bet 1: pays $\$1$ if A and B are true, nothing otherwise.
Price: $\$0.40$ $(= \$1 \times cr(A\&B))$
Bet 2: pays $\$0.40$ if B is false ($\neg B$ is true), nothing otherwise.
Price: $\$0.08$ $(= \$0.40 \times cr(\neg B))$

The agent pays a total of $\$0.48$ for these two bets. If B is false, they win back $\$0.40$, incurring a net loss of $\$0.08$. If the agent learns that B is true (and nothing more), the bookie buys back a third bet from the agent:

Bet 3: pays $\$1$ if A is true, nothing otherwise.
Price: $\$0.40$ $(= \$1 \times cr_B(A))$

Now, with B being true, the agent is also stuck with a loss of $\$0.08$, regardless of whether A is true or false.[7]

In the example above, if the agent planned to update to a credence that is higher than the one warranted by conditionalization, the bookie would have to adjust which bets are bought and which ones are sold. Generally, the argument only works for known updating plans, because they are needed to determine which bets will generate a sure loss. A Dutch strategy cannot be made against agents whose updating plans are unknown. Furthermore, if the agent plans to obey conditionalization, but fails to do so, they are also immune to Dutch strategies. Dutch strategy arguments are thus often viewed as supporting a synchronic norm constraining credences and updating plans[8], which for our purposes can be expressed as:

Rule Conditionalization: If, upon learning some evidence A with certainty (and no other evidence), the agent plans to update their credence in claim B to $cr_A(B)$, then their planned credence equals their current credence in B conditional on A, so $cr_A(B) = cr(B|A)$, assuming that $cr(A) > 0$.

When it is clear from context, we just write "conditionalization". In standard presentations of the Dutch strategy argument, it is assumed that the agent starts out with coherent credences, and then different updating plans are compared to check whether they lead to a sure loss. The reason for this is that if the agent has non-probabilistic initial credences or violates the ratio formula, then there is a Dutch strategy inflicting

---

[7]A well-presented general recipe for constructing a Dutch strategy can be found in Briggs (2009).

[8]A detailed discussion which also includes non-deterministic updating plans can be found in Pettigrew (2020).

sure loss no matter which updating plans the agent adopts. For instance, in example 1, both Jane and Clara are exposed to Dutch strategies. Yet, as we will show shortly, we can devise Dutch strategies against initially incoherent agents that can distinguish between initial incoherence and incoherence generated by updating plans, which will help us identify good updating plans for incoherent agents.

Why do we want to identify a good property of updates to credence assignments that allows for incoherent starting points? What's wrong with having a norm that only works for coherent starting points? Probabilistic coherence of credence assignments is usually seen as a rational ideal, but in practice, it is often violated. Empirical studies in psychology and cognitive science have shown, in a variety of different settings, that human reasoners tend to have incoherent credences, with violations of probabilism being sometimes slight and sometimes more severe.[9] In settings in which an initial credence assignment is a dataset resulting from some sort of judgment aggregation, the collected probability estimates may not be probabilistically coherent (see e.g. Wang, Kulkarni, Poor, and Osherson (2011)). It would be desirable to know whether there is a target property that updates to such incoherent credence assignments should have. In fact, it would be very surprising if there were no way to characterize desirable updates to incoherent credences - after all, our best psychological evidence suggests that human reasoners update their incoherent credences all the time, without their credences spiraling uncontrollably into varying degrees of incoherence. While we're not in search of the psychological mechanism for these updates, the existence of such a mechanism makes us hopeful that we can identify a property that characterizes desirable updates to incoherent credence functions.

It might be suggested that instead of updating incoherent credences directly, one must coherentize incoherent credence assignments and then subsequently update these revised credences. This suggestion addresses how a reasoner should actually go about updating their credences, in an implementation-focused sense. Our proposal has a slightly different aim. It is not intended as directly followable advice for incoherent thinkers who seek to update their credences on new information. Rather, we aim to identify a property that distinguishes desirable from undesirable updates, regardless of whether the starting credences were coherent or incoherent. By what process the agent arrives at their update is not relevant in this context. In trying to identify this property, we take ourselves to be engaging in the same project as other Bayesian epistemologists and decision theorists, who aim to specify what conclusions thinkers should reach, or what decisions they should make, but not how they should get there.We will have reached our

---

[9]Some well-studied coherence violations include the base rate fallacy (see e.g. Kahneman and Tversky (1973), Koehler (1996)), the unpacking effect (see e.g. Tversky and Koehler (1994), Van Boven and Epley (2003)), the conjunction fallacy (see e.g. Tversky and Kahneman (1983) and Crupi, Fitelson, and Tentori (2007)), and violations of the ratio formula (see e.g. Zhao, Shah, and Osherson (2009) and Costello and Watts (2016)).

aim if we can characterize such a property of updates, and justify why updates with this property are desirable.

3 Dutch Books and Degrees of Incoherence

3.1 Distance-based Incoherence Measures and Dutch Books

Standard theories in Bayesian epistemology only distinguish between coherent and incoherent credence assignments and updating strategies, and they tend to have little to say about the incoherent cases. Yet, as De Bona and Staffel have argued, this is an unfortunate oversight (De Bona & Staffel, 2017, 2018; Staffel, 2019). Probabilistic coherence tends to be advertised by Bayesians as a rational ideal that is desirable for non-ideal thinkers to approximate. But in order for coherence to play this role, we need to know more about incoherent credences – what does it mean for them to approximate coherence? And why is it desirable for them to do so?

De Bona and Staffel have proposed that a good strategy to answer these questions is to first identify some valuable property that fully coherent credences have. Then, the task is to show that there is some way of measuring the distance of credence functions to coherence, such that decreasing the distance to coherence delivers increasing portions of this value.[10] If we apply this reasoning to the idea that coherent credences are valuable because they are action guiding, we get the following picture: Let cr be an agent's (unconditional) credences, defined over a set of sentences. (Using sentences rather than propositions has the benefit of allowing agents to assign different credences to logically equivalent claims.) If cr is coherent, then it is not vulnerable to a guaranteed loss from a Dutch book, and hence delivers non-self undermining action guidance. If cr is incoherent, then it is vulnerable to a Dutch book. We can measure both how incoherent cr is and how large of a Dutch book loss it is vulnerable to in a way that correlates the two.

To characterize incoherence measures, we assume the agent has an unconditional credence assignment $cr : \{A_1, \ldots, A_n\} \to [0, 1]$ over a set of sentences A_i from a propositional language $\mathcal{L}$. This assignment can be represented as a set $\{cr(A_1) = q_1, \ldots, cr(A_n) = q_n\}$, with $q_i \in [0, 1]$. In order to measure the distance from cr to coherence, one can measure the distance from the vector $\langle cr(A_1), \ldots, cr(A_n) \rangle \in [0, 1]^n$ to the closest vector $\langle cr^*(A_1), \ldots, cr^*(A_n) \rangle \in [0, 1]$ corresponding to a probabilistic credence assignment cr^*. This distance can be measured with different functions $d : [0, 1]^n \times [0, 1]^n \to [0, \infty)$ yielding different incoherence measures:

$$\mathcal{I}_d(c) = \min\{d(c, c^*) | c^* \text{ is coherent}\}$$

---

[10]Here and below, we sometimes use "distance" in a general way to refer to both distances and divergences in the mathematical sense. Technically, distances, but not divergences, must satisfy symmetry and the triangle inequality.

Common options for the function d include the p-norm distances d_p, such as the Manhattan ($p = 1$), the Euclidean ($p = 2$) and the Chebyshev ($p = \infty$) distances (Thimm, 2013; Potyka, 2014; De Bona & Finger, 2015). For each such distance, $d_p(X, Y)$ is defined as the p-norm of $X - Y$; for instance, the Manhattan (or absolute) distance between two vectors $X, Y \in \mathbb{R}^n$ is given by $d_1(X, Y) = \sum_{i=1}^{n} |X_i - Y_i|$.

As mentioned above, we want to be able to capture the loss in value that comes with increasing incoherence, and we can do so by tracking the size of the sure loss an agent is exposed to via Dutch books. The guaranteed net loss caused by a Dutch book against an incoherent agent can always be arbitrarily scaled up if the bet stakes are not limited somehow. To measure the incoherence through this loss we need to normalize it (Schervish et al., 2002, 2003). Staffel (2019) argues that normalizing Dutch book loss by requiring that each bet has a maximal stake of 1 yields a desirable way of tracking Dutch book loss induced by incoherence. The resulting measure is equivalent to $\mathcal{I}_{d_1}$ for unconditional credences (De Bona & Finger, 2015). In other words, the maximum sure loss an agent is exposed to via a Dutch book with stakes no greater than 1 is equal to the minimum Manhattan distance from their credences to coherent ones. This shows how approximations to coherence can deliver increasing portions of value, the value here being a decreasing vulnerability to losses from standardized Dutch books (we discuss the use of alternative incoherence measures in section 6).

3.2 Augmenting One's Credences without Increasing Incoherence

This delivers a plausible story of why coherence is an ideal worth approximating, since it tells us both what is meant by approximation, and what value is attained by approximating coherence. We can now use this idea and apply it to various cases in which we want to assess incoherent credence assignments. For example, agents often want to draw inferences based on their existing attitudes, such as a very basic modus ponens inference – from *if p then q* and *p*, they infer that *q*. It is fairly well understood how this type of *augmentative reasoning* works for perfectly rational agents. They should simply draw inferences from their existing attitudes that conform to the rules of logic and probability. But can incoherent agents sensibly engage in this kind of reasoning without having to first fix their incoherence? Is there a way to directly assess whether an augmentative inference made by an incoherent agent is an advisable inference for them to make?

As Staffel (2019) shows, the Dutch book based incoherence measure just introduced, combined with the idea that you should not draw inferences that worsen your incoherence, can be used to identify good augmentative inferences whose starting points can be either coherent or incoherent credence assignments. The basic idea is that when an agent fills gaps in their credence assignment based on their existing credences, they should not

become more incoherent than they were to begin with. For coherent agents, this simply means to add new credences that cohere with their existing credences according to the probability axioms. This ensures that the augmented credences will be coherent as well.

For incoherent starting credences, remember that we have singled out an incoherence measure that determines incoherence as the Manhattan distance between the vector representing the agent's incoherent credences and a vector representing some closest coherent credence assignment over the same set of sentences. This measure tracks standardized Dutch book loss if we limit betting stakes to at most 1 per bet. Given this measure, an agent can never become less incoherent by adding an additional credence to her credence assignment, since adding a new credence cannot reduce the distance between her credences and the closest coherent credences. Adding a new credence can only either increase the distance between her credences and some closest coherent credence assignment, or leave it unchanged. Staffel shows that it is always possible to fill gaps in one's existing credence assignment in such a way that it leaves one's degree of incoherence unchanged. Suppose the agent is trying to find the best credence to assign to some claim A_i about which they previously had no opinion. There are two options: either they assign to A_i a credence that coheres with some coherent credence assignment that minimizes the Manhattan distance to their existing credences (call this credence $cr(A_i) = x$), or they assign to A_i a credence that doesn't cohere with any coherent credence assignment that minimizes the Manhattan distance to their existing credences (call this credence $cr(A_i) = y$). If they choose $cr(A_i) = y$, then the distance of the agent's thus augmented credences to some closest coherent credence assignment will be greater than before $cr(A_i) = y$ was added, hence the agent's incoherence and corresponding Dutch book vulnerability are increased. However, it is always possible for the agent to choose some $cr(A_i) = x$ instead. Since $cr(A_i) = x$ coheres with some coherent credence function closest to the agent's existing credences, adding it to the agent's credences will leave the minimum distance between their augmented credences and some closest coherent credences unchanged. It thus represents a conservative choice for the new credence assignment.

Summing up, this shows that the constraint that augmentative inferences should not make agents more incoherent can be applied regardless of whether the agent starts with coherent or incoherent credences. In either case, we can identify which new credence assignments would leave the agent's degree of incoherence unchanged, and are thus desirable augmentative inferences.

Of course, augmentative reasoning is a special case, insofar as it assumes that the agent doesn't change their existing credences and doesn't learn new information. But we're interested in cases in which agents learn new information and plan to update their credences in response. We want to know whether we can identify a property of good *updating plans* that applies regardless of whether the starting point is a coherent or an incoherent credence assignment. It turns out that we can do so by using a similar idea to

the one Staffel employed to identify good augmentative inferences. The basic idea is that an agent's updating plans should not worsen the incoherence already contained in their starting credences. To implement this idea, we need to generalize our distance-based Dutch book measure to make sure it can apply to updating plans, and explain how it tracks the Dutch book loss from Dutch strategies. Applying this idea to coherent agents yields the result that only rule conditionalization avoids plan incoherence. For incoherent agents, we can similarly identify updates that keep their plan incoherence at zero. We will explain the details of this proposal in the next section.

4 Plan Incoherence and Tolerant Conditionalization

4.1 Measuring Dutch Strategy Loss

Typical Dutch strategy arguments cannot distinguish between two different updating strategies if the agent is already initially incoherent, since they will be exposed to sure loss regardless of their updating plans. This sure loss would come from the regular Dutch book the agent is vulnerable to for possessing incoherent initial credences. Nevertheless, if the amount to be gambled in a Dutch strategy is normalized in some way, some updating strategies can yield higher sure loss than others. To quantify this difference, we can measure the sure loss inflicted by a Dutch strategy normalized by the greatest stake involved, which is equivalent to measuring guaranteed loss when the stakes are no greater than 1.

A Dutch strategy differs from a Dutch book due to the possibility of making bets at two different times: initially, at t_1; and at t_2, after the agent learns the truth of some element E from a given partition. The bets to be placed at t_2 are a function of what is learned. Suppose a Dutch strategy involves a bet on A at t_2 if exactly E is learned. If this bet pays λ (the stake), the fair price to the agent would be $\lambda \times cr_E(A)$, where $cr_E(.)$ denotes the updated credences after the agent learns E. As this bet is to occur only if E is learned, it behaves exactly like a conditional bet on $(A|E)$ at t_1: the net loss is zero if E is false, otherwise it works like a bet on A. Hence, if we want to compute the guaranteed loss to which an agent with updating plans $cr_E(.)$ for a fixed E is vulnerable from a Dutch strategy, we can compute the same loss if we represent their updating plans as conditional credences $cr(.|E)$ instead and set up a synchronic Dutch book against the agent. In fact, this identity holds even if the agent has updating plans for different sentences E, as long as they belong to the partition they learn from. This correspondence reflects the link between Dutch books against agents violating the ratio formula and Dutch strategies against agents violating rule conditionalization. Example 3 illustrates this by giving the synchronic Dutch book that corresponds to the diachronic one from example 2:

Example 3. *Consider the agent from example 2 with an extra conditional credence $cr(A|B) = 0.4$ corresponding to the updating plan $cr_B(A) = 0.4$:*

  $cr(A\&B) = 0.4$, $cr(\neg A\&B) = 0.4$, $cr(B) = 0.8$, $cr(\neg B) = 0.2$ and $cr(A|B) = 0.4$.
  The bookie will sell the agent the same two bets:

  Bet 1: pays \$1 if A and B are true, nothing otherwise.
  Price: \$0.40 (= \$1 × $cr(A\&B)$)
  Bet 2: pays \$0.40 if B is false ($\neg B$ is true), nothing otherwise.
  Price: \$0.08 (= \$0.40 × $cr(\neg B)$)

  When B is false, these bets yield a loss of \$0.08, as before. But now the bookie also buys the following conditional bet, which has an effect only if B is true:

  Bet 3: pays \$1 if A is true, nothing otherwise. The bet is called off if B is false.
  Price: \$0.40 (= \$1 × $cr(A|B)$)

  The conditional gamble on $(A|B)$ works like a bet on A when B is true. Thus, if B is true, the loss is also \$0.08, as in example 2.

To formalize losses from Dutch strategies, assume a partition $\mathcal{P} \subseteq \mathcal{L}$ [11] from which the agent is to learn a proposition. We also assume a set of pairs $(A_i, |A_j) \in \mathcal{L} \times \mathcal{L}$ to which the agent assigns initial credences $cr(A_i|A_j)$ and a set of pairs $(A_l, E_k) \in \mathcal{L} \times \mathcal{P}$ for the updating plans $cr(A_l|E_k)$. Call a *credal strategy* a set containing credences $cr(A_i|A_j) = q_{ij}$ [12] and updating plans $cr_{E_k}(A_l) = q_{kl}$ for arbitrary $q_{ij}, q_{kl} \in [0,1]$. [13] An updating plan $cr_E(A) = q$ can be construed as a disposition to update the credence in A to q if exactly E is learned with certainty. The credences and plans can be over different propositions, for instance: $\Psi = \{cr(A) = 0.1, cr_C(B) = 1, cr_E(D) = 0\}$, for $C, E \in \mathcal{P}$. Note that unconditional credence assignments can be viewed as credal strategies with no updating plans.

As standardly defined, conditionalization refers to an agent's conditional credences. Yet, we've defined credal strategies as being possibly gappy, in that they may lack the conditional credences corresponding to the agent's updating plans. To deal with this, we propose to extend the definition of rule conditionalization. An agent who lacks the relevant conditional credences still counts as obeying rule conditionalization as long as there is a probability function pr^* extending the initial credences in the agent's credal strategy Γ such that, for each of their updating plans $cr_E(A) = q \in \Gamma$, it holds that

---

[11]We say a set of sentences forms a partition if they are pairwise inconsistent and their disjunction is valid.

[12]We use $cr(A_i|\top)$ to represent the unconditional credence $cr(A_i)$.

[13]We use the term "unconditional credal strategy" to indicate that we are not yet incorporating agents' conditional credences, only their updating plans, which need not be the same. While we can represent updating plans as conditional credences for technical purposes as explained above, we don't assume they must be identical in real agents. See section 5 and also Zhao, Crupi, Tentori, Fitelson, and Osherson (2012).

$q \times pr^*(E) = pr^*(A\&E)$. This understanding of rule conditionalization still assumes that the agent obeys probabilism and the ratio formula.

For any unconditional credal strategy Γ, let Γ^{sync} denote the credal strategy formed by replacing each updating plan $cr_E(A) = q$ by a conditional credence $cr(A|E) = q$; for instance, $\Psi^{sync} = \{cr(A) = 0.1, cr(B|C) = 1, cr(D|E) = 0\}$. If stakes are limited in the same way, the maximum guaranteed loss an agent Γ is exposed to via a Dutch strategy is equal to the maximum amount the agent Γ^{sync} can surely lose in a Dutch book.[14]

We saw in section 3.1 that, for unconditional credences, the maximum sure loss from Dutch books an agent is exposed to when the stakes are no greater than 1 equals the Manhattan distance from her credences to some closest coherent credences. This result can be extended to conditional credences and conditional gambles, via violation incoherence measures (Potyka, 2014).

Given a probability measure $pr : \mathcal{L} \to [0,1]$ over the whole language, each conditional credence $cr(A_i|E_i) = q_i$ corresponds to a *violation* ε_i defined as:

$$\varepsilon_i = pr(A_i\&E_i) - q_i \times pr(E_i)$$

Note that $\varepsilon_i = 0$ iff $pr(A_i\&E_i) = q_i \times pr(E_i)$. For an unconditional credence $cr(A_i) = q_i$, its violation is defined via $cr(A_i|\top) = q_i$.

Suppose $\Gamma = \{cr(A_1|E_1) = q_1, \ldots, cr(A_n|E_n) = q_n\}$ is a set of conditional and unconditional credences. Each probability measure pr yields a vector of violations $\langle \varepsilon_1, \ldots, \varepsilon_n \rangle$. Their elements can all be zero iff the credences are probabilistic and satisfy the ratio formula. We can measure the incoherence of a set of credences as the minimum of the 1-norm of this violation vector. Formally, the incoherence measure $\mathcal{I}_\varepsilon$ can be defined as:

$$\mathcal{I}_\varepsilon(\Gamma) = \min_{pr} ||\langle \varepsilon_1, \ldots, \varepsilon_n \rangle||_1 = \min_{pr} \sum_{i=1}^{n} |\varepsilon_i| = \min_{pr} \sum_{i=1}^{n} |pr(A_i\&E_i) - q_i \times pr(E_i)|$$

.

Since $pr(\top) = 1$ and $pr(A_i\&\top) = pr(A_i)$ for any probability measure pr, $cr(A_i) = q_i$ yields the violation $\varepsilon_i = pr(A_i) - q_i$. Consequently, for a set of unconditional credences Γ, it holds that $\mathcal{I}_{d_1}(\Gamma) = \mathcal{I}_\varepsilon(\Gamma)$. Moreover, if Γ also contains conditional credences, $\mathcal{I}_\varepsilon(\Gamma)$ is the maximum sure loss the agent Γ is exposed to via a Dutch book whose stakes are no greater than 1, possibly including conditional gambles (De Bona & Finger, 2015).

Putting the pieces together, the maximum sure loss from a Dutch strategy that an agent with credal strategy Γ is exposed to, when the stakes are no greater than 1, is equal to $\mathcal{I}_\varepsilon(\Gamma^{sync})$. We can demonstrate this way of measuring the Dutch book loss of a credal strategy by applying it to the agent from example 2.

---

[14]We're abusing the notation here, equating agents with their credal strategies.

Example 4. *In example 2, the agent's credal strategy is:*

$$\Gamma = \{cr(A\&B) = 0.4, cr(\neg A\&B) = 0.4, cr(B) = 0.8, cr(\neg B) = 0.2, cr_B(A) = 0.4\}$$

To determine the maximum sure loss from a Dutch strategy with stakes no greater than 1, we can apply $\mathcal{I}_\varepsilon$ to the credal strategy with updating plans replaced by conditional credences:

$$\Gamma^{sync} = \{cr(A\&B) = 0.4, cr(\neg A\&B) = 0.4, cr(B) = 0.8, cr(\neg B) = 0.2, cr(A|B) = 0.4\}$$

The sure loss is then given by $\mathcal{I}_\varepsilon(\Gamma^{sync})$, which can be computed by finding a probability measure pr solving the optimization problem below:

$$\min |\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3| + |\varepsilon_4| + |\varepsilon_5|$$
$$pr(A\&B) - 0.4 = \varepsilon_1$$
$$pr(\neg A\&B) - 0.4 = \varepsilon_2$$
$$pr(B) - 0.8 = \varepsilon_3$$
$$pr(\neg B) - 0.2 = \varepsilon_4$$
$$pr(A\&B) - 0.4 \times pr(B) = \varepsilon_5$$

The solution pr^* is such that $pr^*(A\&B) = pr^*(\neg A\&B) = 0.4$ and $pr^*(A|B) = 0.5$, yielding $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = 0$ and $|\varepsilon_5| = 0.08$. The resulting $\mathcal{I}_\varepsilon(\Gamma^{sync}) = 0.08$ is the guaranteed loss the agent with credal strategy Γ would be vulnerable to via a Dutch strategy, limiting the highest stake to 1. This is the same sure loss an agent with credal strategy Γ^{sync} is exposed to via a Dutch book with conditional gambles.

4.2 Plan Incoherence and Tolerant Conditionalization

The maximum sure loss a Dutch strategy can extract from an agent's credal strategy, which includes their initial credences and updating plans, cannot be smaller than the maximum sure loss the agent is vulnerable to from a Dutch book on their initial credences alone (with stakes no greater than 1 per bet). Different sets of updating plans, yielding different credal strategies, can result in a bigger or smaller increase to this sure loss. We call this increase *plan incoherence*.

Credal strategies Γ can be partitioned into two parts: initial credences Γ_{cr} and updating plans Γ_{plan}. Now we can formalize the *plan incoherence* of an agent with credal strategy Γ as $\mathcal{I}_\varepsilon(\Gamma^{sync}) - \mathcal{I}_\varepsilon(\Gamma_{cr})$, as illustrated in example 5.

Example 5. *Recall the agent from examples 4 and 2, whose initial credences are:*

$$\Gamma_{cr} = \{cr(A\&B) = 0.4, cr(\neg A\&B) = 0.4, cr(B) = 0.8, cr(\neg B) = 0.2\}$$

Their updating plan is $\Gamma_{plan} = \{cr_B(A) = 0.4\}$, forming the credal strategy $\Gamma = \Gamma_{cr} \cup \Gamma_{plan}$.

As the initial credences are probabilistic, $\mathcal{I}_{d_1}(\Gamma_{cr}) = \mathcal{I}_\varepsilon(\Gamma_{cr}) = 0$. But from example 4 we know that $\mathcal{I}_\varepsilon(\Gamma^{sync}) = 0.08$, implying a plan incoherence for the credal strategy Γ of $\mathcal{I}_\varepsilon(\Gamma^{sync}) - \mathcal{I}_\varepsilon(\Gamma_{cr}) = 0.08 - 0 = 0.08$.

The plan incoherence can never be negative, as the Dutch book for the initial credences can be replicated in the Dutch strategy for the credal strategy. We have not yet shown if and when the plan incoherence can be zero for arbitrary initial credences. The incoherence of the initial credences $\mathcal{I}_\varepsilon(\Gamma_{cr})$ is computed via a probability measure pr^* minimizing the 1-norm of the violations. For unconditional credences cr, this pr^* is an extension of a Manhattan-closest coherent credence assignment cr^*, which minimizes $d_1(cr, cr^*)$. If the conditional credences in Γ_{plan}^{sync} yield null violations for such a pr^*, the plan incoherence is zero. This is achieved if the updating plans $cr_E(A) = q \in \Gamma_{plan}$ are such that $q = pr^*(A|E)$ whenever $pr^*(E) > 0$. Hence, for any set of initial credences Γ_{cr}, there is always a way of extending it with updating plans Γ_{plan}, forming $\Gamma = \Gamma_{cr} \cup \Gamma_{plan}$ such that $\mathcal{I}_\varepsilon(\Gamma^{sync}) = \mathcal{I}_\varepsilon(\Gamma_{cr})$ and the plan incoherence is equal to zero. Actually, this is the only way to achieve null plan incoherence (see Corollary 1 in the Appendix).

In particular, if Γ_{cr} is coherent, it can be extended to a probability measure pr^*, and zero plan incoherence corresponds to conditionalizing pr^* – i.e., updating plans in Γ_{plan} have the form $cr_E(A) = pr^*(A|E)$, whenever $pr^*(E) > 0$, for a pr^* extending cr.

Hence, an agent with possibly incoherent credences always has two options: adopting updating plans that increase the sure loss they are exposed to from a Dutch strategy with stakes lesser or equal to 1 or adopting updating plans that keep this loss constant. The latter option is captured by the following property:

Tolerant Conditionalization: The agent's credal strategy has zero plan incoherence.

For an agent with probabilistic initial credences obeying the ratio formula, positive plan incoherence is equivalent to Dutch strategy vulnerability, and tolerant conditionalization is equivalent to rule conditionalization. For agents with incoherent initial credences, we can see the difference between violating and satisfying TC in our agents from example 1:

Example 6. *Recall Jane and Clara from example 1. They have the same initial credences:*

$$J_{cr} = C_{cr} = \{cr(A\&B) = 0.4, cr(\neg A\&B) = 0.4, cr(B) = 0.9, cr(\neg B) = 0.1\}$$

Jane's updating plan is $J_{plan} = \{cr_B(A) = 0.5\}$, while Clara's is $C_{plan} = \{cr_B(A) = 0.3\}$. Their credal strategies are then $J = J_{cr} \cup J_{plan}$ and $C = C_{cr} \cup C_{plan}$, respectively.

As both have incoherent initial credences, both are vulnerable to sure loss from Dutch books. The bookie can buy bets with \$1 stakes on $A\&B$ and $\neg A\&B$, each for \$0.4, and sell a bet paying \$1 if B is true for \$0.9, thus inflicting a net loss of \$0.1 on each of them. The maximum guaranteed loss when stakes are normalized is thus $\mathcal{I}_\varepsilon(J_{cr}) = \mathcal{I}_\varepsilon(C_{cr}) = 0.1$.

The maximum sure loss above reflects the Manhattan distance from the initial credences to a closest coherent credence assignment cr^*, for instance with $cr^*(A\&B) = cr^*(\neg A\&B) = 0.45$, yielding $cr^*(B) = 0.9$ and $cr^*(\neg B) = 0.1$. By the ratio formula, these credences entail $cr^*(A|B) = 0.5$, which matches Jane's plan J_{plan}. Hence, adding this conditional credence to the initial credences, forming J^{sync}, does not increase the Manhattan distance to coherence, and $\mathcal{I}_\varepsilon(J^{sync}) = \mathcal{I}_\varepsilon(J_{cr})$. This implies that Jane's plan incoherence, $\mathcal{I}_\varepsilon(J^{sync}) - \mathcal{I}_\varepsilon(J_{cr})$, is equal to zero, and tolerant conditionalization holds.

To check whether Clara also obeys tolerant conditionalization, we need to compute $\mathcal{I}_\varepsilon(C^{sync})$, which is equal to the maximum sure loss she is exposed to via a Dutch strategy when stakes are normalized, and compare this to her initial incoherence. Recall that $C^{sync} = C_{cr} \cup C^{sync}_{plan}$, where $C^{sync}_{plan} = \{cr(A|B) = 0.3\}$, so $\mathcal{I}_\varepsilon(C^{sync})$ can be determined by finding a probability measure pr solving the optimization problem below:

$$\min |\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3| + |\varepsilon_4| + |\varepsilon_5|$$
$$pr(A\&B) - 0.4 = \varepsilon_1$$
$$pr(\neg A\&B) - 0.4 = \varepsilon_2$$
$$pr(B) - 0.9 = \varepsilon_3$$
$$pr(\neg B) - 0.1 = \varepsilon_4$$
$$pr(A\&B) - 0.3 \times pr(B) = \varepsilon_5$$

One possible solution pr^* is such that $pr^*(A\&B) = 0.4$, $pr^*(\neg A\&B) = 0.5$ and $pr^*(A|B) = 4/9$, yielding $\varepsilon_1 = \varepsilon_3 = \varepsilon_4 = 0$, $\varepsilon_2 = 0.1$ and $\varepsilon_5 = 0.13$. Thus, $\mathcal{I}_\varepsilon(C^{sync}) = 0.23$, and the plan incoherence is $\mathcal{I}_\varepsilon(C^{sync}) - \mathcal{I}_\varepsilon(C_{cr}) = 0.13$. We can conclude that Clara violates tolerant conditionalization.

The pr^* above indicates that the updating plan $cr_B(A) = 4/9$, together with the initial credences C_{cr}, would satisfy tolerant conditionalization. In fact, any $cr_B(A) = q$ with $q \in [4/9, 5/9]$ would do so, as there would be a probability measure pr^*, with $pr^*(A|B) = q$, extending a credence assignment cr^* that is Manhattan-closest to cr.

We can see now that TC is similar in spirit to the strategy for augmentative reasoning from section 3. The basic idea in both cases is that we identify a closest coherent credence assignment cr^* to the agent's existing incoherent credences, which represents a conservative coherentization of the agent's credences. Then we identify either a new unconditional credence (in the augmentative case) or an updating plan that coheres with cr^*. This method ensures that neither the augmented credence nor the updating plan

worsen the agent's existing incoherence. However, there is also an important difference between the two cases. Agents who augment their credences in the way we've described don't worsen their incoherence, but they also don't become coherent. By contrast, an agent who obeys tolerant conditionalization will end up with coherent credences if they follow their updating plans. We think this is a sensible outcome, since TC is supposed to identify the most desirable update that is available to an incoherent agent. However, the fact that TC recommends coherent credences of course doesn't mean that incoherent agents will in fact become coherent upon updating their credences. We can interpret varying degrees of plan incoherence as identifying better and worse updating plans agents might adopt.

The cases discussed so far concern agents who learn new information, but we can also consider what happens when an agent "learns" a tautology. For coherent agents, conditionalizing on a tautology leaves their credences unchanged. By contrast, when an incoherent agent considers updating on a tautology, tolerant conditionalization is satisfied just in case the agent plans to move to a Manhattan-closest coherent credence assignment. TC can thus also act as a property of coherentizing updates in the absence of new empirical information. This seems like a desirable property, since it aligns with plausible existing suggestions for how incoherent agents should coherentize their attitudes.

While there is a sense in which tolerant conditionalization is a more general version of rule conditionalization, there is also an important difference between them. Rule conditionalization is usually understood as a norm that rational agents should obey whenever they are to learn some new information from a partition with certainty. We think that this characterization as a norm is too strong for tolerant conditionalization. Instead, TC should be understood as identifying updating plans that have specific pragmatic benefits. Every probabilistic, conditionalizing credal strategy has these benefits, as well as some credal strategies with non-probabilistic initial credences. In specific cases, we can then ask whether these benefits provide a good enough reason to favor an update that complies with TC, or if there are overriding considerations that call for a different update. This could happen, for example, if an incoherent agent is somehow aware that a particular coherentization of their attitudes is rationally called for, even though it violates TC. TC, and more generally the idea that agents should try to minimize increases in incoherence in their reasoning, are based on the thought that incoherent agents who have no insight into the origin of their incoherence should adjust their credences in a conservative manner. However, if an agent finds out somehow that a particular type of error is the source of their incoherence, then they should fix this particular mistake even if their credence change doesn't align with TC. In such a case, the benefits of complying with TC are outweighed.

5 Incorporating Conditional Credences

So far, our representations of agents' credence assignments have only incorporated their unconditional credences and their updating plans. These simplified representations didn't incorporate their conditional credences. A conditional credence $cr(A|B) = q$ can be interpreted as the agent's credence q in A on the supposition that B is true, which is closely related to, but not identical with an updating plan for learning B. For coherent agents, we know that their conditional credences relate to their unconditional credences via the ratio formula, and that their updating plans match their conditional credences, but we cannot assume this is true for incoherent agents. Zhao et al. (2009) provide evidence that human reasoners' unconditional and conditional credences are not related via the ratio formula. Zhao et al. (2012) found that in human reasoners, the conditional credence of an event A supposing an event B does not match the credence of A after having learned B. It is desirable to be able to represent these ways of being incoherent in our models. In what follows, we show how to incorporate conditional credences into our representations of incoherent agents' credences.

An agent possessing both unconditional and conditional credences can be incoherent in two ways: the unconditional credences might be non-probabilistic or the conditional credences might violate the ratio formula. In either case, the agent is vulnerable to a Dutch book, and the corresponding sure loss, when stakes are no greater than 1, can be measured via $\mathcal{I}_\varepsilon$. We can again define plan incoherence as the increase in sure loss caused by considering the updating plans as additional conditional credences. Tolerant conditionalization then still requires null plan incoherence. For coherent initial credences, updating plans equal to the conditional credences keep the agent immune to Dutch books, thus tolerant conditionalization is equivalent to conditionalization. When the agent is initially incoherent, tolerant conditionalization may or may not demand that the updating plans match the conditional credences, as we shall see.

The sure loss caused by a Dutch strategy can be measured as the sure loss from a Dutch book if we represent updating plans as conditional credences, as shown in Section 4.1. This approach can be extended to credence functions that include initial conditional credences, but we need to handle possible duplicated credences in Γ^{sync} coming from the updating plans. For instance, if $cr(A|B) = q \in \Gamma_{cr}$ is a conditional credence and $cr_B(A) = q' \in \Gamma_{plan}$ is an updating plan, we need a way to consider both at the same time in the set Γ^{sync}, as different elements, to compute the Dutch book sure loss corresponding to the Dutch strategy loss. In this case, $cr_B(A) = q' \in \Gamma_{plan}$ is replaced by $cr(A'|B) = q'$ in Γ^{sync}_{plan}, where A' is logically equivalent to A (say $A\&A$) but $cr(A'|B) = q'$ is not already in the initial credences Γ_{cr}.

Example 7. *Recall Jane from examples 1 and 6, with an additional conditional credence*

$cr(A|B) = 0.5$:

$$J_{cr} = \{cr(A\&B) = 0.4, cr(\neg A\&B) = 0.4, cr(B) = 0.9, cr(\neg B) = 0.1, cr(A|B) = 0.5\}$$

Jane's updating plan is $J_{plan} = \{cr_B(A) = 0.5\}$, and her credal strategy is $J = J_{cr} \cup J_{plan}$.

In example 6, the updating plan $cr_B(A) = 0.5$ did not increase her incoherence when considered together with her unconditional credences. Consequently, neither does her initial conditional credence $cr(A|B) = 0.5$, and her initial incoherence remains $\mathcal{I}_\varepsilon(J_{cr}) = 0.1$. Any probability measure pr^* minimizing the 1-norm of the violations in J_{cr} has $pr^*(A|B) = 0.5$, so considering her updating plan as an additional conditional credence, $J_{plan}^{sync} = \{cr(A'|B) = 0.5\}$, does not increase her initial incoherence: $\mathcal{I}_\varepsilon(J^{sync}) = \mathcal{I}_\varepsilon(J_{cr} \cup J_{plan}^{sync}) = 0.1$. Therefore, her plan incoherence is $\mathcal{I}_\varepsilon(J^{sync}) - \mathcal{I}_\varepsilon(J_{cr}) = 0$, and tolerant conditionalization is satisfied.

In the example above, tolerant conditionalization is satisfied by updating plans that are identical to the initial conditional credences, as standard conditionalization would require, even though the latter violate the ratio formula. However, this is not generally true. Matching updating plans with initial conditional credences is neither a sufficient nor a necessary condition for satisfying TC, as the next example shows.

Example 8. *Recall Clara from examples 1 and 6, with an additional conditional credence* $cr(A|B) = 0.3$:

$$C_{cr} = \{cr(A\&B) = 0.4, cr(\neg A\&B) = 0.4, cr(B) = 0.9, cr(\neg B) = 0.1, cr(A|B) = 0.3\}$$

Clara's updating plan is $C_{plan} = \{cr_B(A) = 0.3\}$, and her credal strategy is $C = C_{cr} \cup C_{plan}$.

Note that C^{sync} from example 6 is identical to C_{cr} here, so that we have $\mathcal{I}_\varepsilon(C_{cr}) = 0.23$. Any probability measure pr^* minimizing the 1-norm of the violations in C_{cr} has $pr^*(A|B) \neq 0.3$. Hence, considering her updating plan as an additional conditional credence, $C_{plan}^{sync} = \{cr(A'|B) = 0.3\}$, does increase the incoherence, namely to $\mathcal{I}_\varepsilon(C^{sync}) = \mathcal{I}_\varepsilon(C_{cr} \cup C_{plan}^{sync}) = 0.36$. Tolerant conditionalization is thus violated, as her plan incoherence is $\mathcal{I}_\varepsilon(C^{sync}) - \mathcal{I}_\varepsilon(C_{cr}) = 0.13$. Even though the updating plan matches her initial conditional credence, it ends up increasing the sure loss the agent is exposed to, for it does not match the closest coherent credences.

What updating plan $cr_B(A) = q$ could Clara alternatively adopt to satisfy tolerant conditionalization, given her initial credences C_{cr}? Recall from example 6 that there's a probability measure pr^* minimizing the 1-norm of the violations in C_{cr} here that yields $pr^*(A|B) = 4/9$. This means that an updating plan $C_{plan} = \{cr_B(A) = 4/9\}$ would imply $\mathcal{I}_\varepsilon(C^{sync}) = \mathcal{I}_\varepsilon(C_{cr}) = 0.26$, with null plan incoherence. In this case, tolerant condition-

alization would be satisfied by an updating plan that differs from the initial conditional credence.

The example above shows that Clara would be exposed to a lesser sure loss via Dutch strategies if her updating plan violated standard conditionalization's assumption that $cr_B(A) = cr(A|B)$. This norm is intended to apply to probabilistic agents satisfying the ratio formula, and we can see that conditional credences violating the ratio formula are not reliably good updating guides. Yet, the converse question remains open: Are conditional credences satisfying the ratio formula always a good updating guide? The norm of conditionalization gives us an affirmative answer in the probabilistic case, but in non-probabilistic scenarios, conditional credences obeying the ratio formula can lead reasoners astray, as the next example shows.

Example 9. *Consider an agent with the following initial incoherent credences:*

$$\Gamma_{cr} = \{cr(A\&B) = 0.4, cr(\neg A\&B) = 0.4, cr(B) = 0.9, cr(\neg B) = 0.2, cr(A|B) = 4/9\}$$

Note that $cr(A|B) = cr(A\&B)/cr(B)$, so the ratio formula is satisfied. Nonetheless, any probability measure pr^* minimizing the 1-norm violation has $pr^*(A\&B) = pr^*(\neg A\&B) = 0.4$, yielding $pr^*(A|B) = 0.5$. This pr^* yields $\mathcal{I}_\varepsilon(\Gamma_{cr}) = 13/90 \approx 0.144$. An updating plan $\Gamma_{plan} = \{cr_B(A) = 0.5\}$ would satisfy tolerant conditionalization, while any other value for $cr_B(A)$, including $cr_B(A) = 4/9$, would increase the sure loss, implying positive plan incoherence.

This means that the agent would be better off with a conditional credence $cr(A|B) = 0.5$, violating the ratio formula, instead of $cr(A|B) = 4/9$. In fact, we would have $\mathcal{I}_\varepsilon(\Gamma_{cr}) = 0.1$ in that case.

The example above reflects the fact that Dutch book arguments for the ratio formula, just like those for conditionalization, rely on probabilistic unconditional credences. For agents violating probabilism, it may be better to violate the ratio formula and/or standard conditionalization, given the aim of minimizing incoherence (as measured by sure loss via Dutch books when stakes have a fixed upper bound).

6 Alternative Incoherence Measures and Relations to Accuracy

6.1 Alternative Incoherence Measures

Tolerant conditionalization can be read as requiring the agent not to increase their incoherence when extending their initial credences with conditional credences corresponding to their updating plans. This correspondence was supported by the equivalence between

Dutch strategies and Dutch books, explained in section 4.1. We've defined incoherence measures via the maximum guaranteed loss an agent is exposed to from a Dutch book containing suitably bounded gambles. To achieve this, we have been limiting the maximum bet prize or stake, which results in an incoherence measure equivalent to the Manhattan distance to the closest coherent credences, for unconditional initial credences. There are additional ways of normalizing the loss from a Dutch book, and the corresponding incoherence measures can be employed to derive variations of tolerant conditionalization.

Schervish, Seidenfeld and Kadane (2002; 2003) have put forward a whole family of incoherence measures based on guaranteed Dutch book loss. They suggest that the bets in a Dutch book are limited via a function f of the bet sizes, where f can either consider only the amount (called the *escrow*) the agent or the bettor can lose in each bet, or their sum, which is the stake. The function f must be continuous, non-decreasing in each argument and invariant under permutation of its arguments, among other properties. These properties define a whole spectrum of functions within two particular extreme cases, the maximum and the sum, on which we focus. In total, six particular ways of normalizing the sure loss from sets of bets stand out, as one can consider the agent's escrows (a), the bettor's escrows (b) or the total stake (s) of the bets, and then bound either their maximum or their sum: $\mathcal{I}_{max}^a$, $\mathcal{I}_{sum}^a$, $\mathcal{I}_{max}^b$, $\mathcal{I}_{sum}^b$, $\mathcal{I}_{max}^s$ and $\mathcal{I}_{sum}^s$.

We employed $\mathcal{I}_{max}^s = \mathcal{I}_\varepsilon$ in formulating TC, because it has some desirable properties that its alternatives lack, as Staffel (2019) explains in some detail. Focusing on the stakes of a bet, instead of the agent's or bookie's escrow, has the advantage of being able to capture the degree of incoherence resulting from credences in tautologies and contradictions. By contrast, escrow-based measures have trouble capturing this and are best used only for credences in contingent claims. We furthermore prefer the max- to the sum-normalization for collections of bets. This is because the max-normalization tracks incoherence anywhere in an agent's credence assignment, whereas the sum-normalization tracks only worst-case incoherence. For example, if an agent assigns zero credence to a tautology, this will entirely determine the agent's degree of incoherence, regardless of what their other credences look like, if we use the sum-normalization. The max-normalization does not have this feature.

Another reason to adopt $\mathcal{I}_{max}^s$ comes from the possibility of characterizing TC when credences are viewed as probability lower/upper bounds. When the agent assign credences over a partition or an algebra, and these assignments can be seen as a set of consistent probability lower/upper bounds, TC is equivalent to conditionalizing a probability measure satisfying those bounds (see Theorem 2 and Corollary 2 in the Appendix). This means that an underconfident (or overconfident) agent that assigns numeric credences cr as if they were probability lower (upper) bounds, defining a set of probability measures, should conditionalize a measure in this set to satisfy TC. The result also applies to Dempster-Shafer belief/plausibility functions since they can be seen as consistent

probability lower/upper bounds.

Even though we prefer $\mathcal{I}_{max}^s$, we don't want to preclude readers from using their preferred incoherence measure, and defining alternative versions of TC using measures of their choice. Any incoherence measure $\mathcal{I}$ based on Dutch book loss, with its particular way of limiting the bets, can be used to gauge the increase in incoherence caused by the updating plans in a credal strategy Γ, defined as $\mathcal{I}(\Gamma^{sync}) - \mathcal{I}(\Gamma_{cr})$, which we call *plan $\mathcal{I}$-incoherence*. TC can then be parametrized by a measure $\mathcal{I}$, demanding the plan $\mathcal{I}$-incoherence to be null, when the original TC would be a particular instance, with $\mathcal{I} = \mathcal{I}_{\varepsilon}$.

$\mathcal{I}$-Tolerant Conditionalization: The agent's credal strategy has zero $\mathcal{I}$-plan incoherence.

Each instance of $\mathcal{I}$-TC has distinctive properties that can be advantageous in particular scenarios. For instance, suppose the agent has incoherent credences $cr(.)$, over a partition or an algebra, that can be made coherent if multiplied by a normalizing factor $0 < \alpha < 1$, such that $cr^*(.) = \alpha \times cr(.)$ is coherent. In this setting, $\mathcal{I}_{sum}^a$-TC is equivalent to conditionalizing $cr^*(.)$; that is, to first normalizing then conditionalizing. When $\alpha > 1$, this equivalence holds for $\mathcal{I}_{sum}^b$-TC instead.

In principle, plan $\mathcal{I}$-incoherence and $\mathcal{I}$-TC can be defined for any incoherence measure[15], not only for those based on Dutch books. Nevertheless, the link between the updating plans and the conditional credences is loosened when incoherence is not measure as Dutch book/strategy loss. Once we stop using Dutch books, it is not obvious that the incoherence of a credal strategy Γ should be measured as the incoherence of Γ^{sync}. Apart from that, plugging different measures into $\mathcal{I}$-TC might be desirable for other reasons, for instance due to benefits in epistemic utility. In any case, we think that versions of TC need to be justified by showing that they track some valuable property of updates.

6.2 Relations to Accuracy

Rule conditionalization has recently been supported by depragmatized arguments, based not on Dutch strategies, but on purely epistemic considerations (Briggs & Pettigrew, 2020; Greaves & Wallace, 2006; Oddie, 1997). Those arguments view accuracy as an epistemic good to be pursued and show how a probabilistic agent who obeys rule conditionalization, rather than another updating rule, has an accuracy advantage. In this section we explore how our notion of plan incoherence can be instantiated to generate a target property for the credal strategies of initially incoherent agents based on accuracy.

---

[15]We assume incoherence measures are monotonic ($\mathcal{I}(\Gamma) \leq \mathcal{I}(\Gamma \cup \Psi)$), otherwise plan incoherence could be negative.

Greaves and Wallace (2006) and Oddie (1997) propose an expected epistemic utility argument in support of rule conditionalization. Their argument relies on an expectation calculated via the agent's initial credences, assumed to be coherent. In a non-probabilistic scenario, it is not clear how this expectation would be computed, which is why we won't try to extend their argument to initially incoherent agents.

A different approach, proposed by Briggs and Pettigrew (2020) and improved by Nielsen (2021), employs the accuracy of the whole credal strategy to ground rule conditionalization in accuracy-dominance terms. For any reasonable accuracy measure, any non-conditionalizing credal strategy is strongly accuracy-dominated by some probabilistic, conditionalizing one (the latter would do better than the former in any possible world) that is not even weakly dominated in this sense. Furthermore, probabilistic, conditionalizing credal strategies are never strongly accuracy-dominated. Note that the dominance is between credal strategies, so the initial credences are not held fixed. As there are no expectations involved, we can try to extend this approach to non-probabilistic initial credences.

We are looking for a target property that can be applied to agents with fixed incoherent initial credences Γ_{cr} to evaluate their updating plans Γ_{plan}. A natural proposal would be to demand that the credal strategy not be accuracy-dominated by another strategy with the same initial credences. That is, given their initially incoherent credences, there would be no other updating plans with which the agent would do better in any possible world. Unfortunately, any probabilistic set of updating plans Γ_{plan} would yield non-dominance in that sense. That is, to dominate any credal strategy with probabilistic updating plans, one would need to change the initial credences as well, not only the updating plans. Probabilistic updating plans are certainly desirable, but we aim for a stronger property that relates them somehow to the agent's initial credences.

If an agent's credal strategy is accuracy-dominated by a probabilistic, conditionalizing one, perhaps they should adopt the updating plans of the latter. However, there might be a set of different dominating credal strategies to pick plans from, and the agent's current plans might in principle already match those in a dominating credal strategy. Still, if all dominating options would change their updating plans, the agent's updating plans seem to be flawed. Intuitively, if the agent is incoherent, and their credal strategy is thus dominated, their updating plans seem to be better if they can be held fixed while moving to a dominating credal strategy. This is captured by the following property, where $\mathfrak{I}$ is an inaccuracy measure that evaluates credal strategies in possible worlds.

$\mathfrak{I}$-*Stability*: Given non-probabilistic credences Γ_{cr}, the updating plans Γ_{plan} are $\mathfrak{I}$-stable if there is a probabilistic, conditionalizing credal strategy $\Psi_{cr} \cup \Gamma_{plan}$ that is more $\mathfrak{I}$-accurate than $\Gamma_{cr} \cup \Gamma_{plan}$ in all possible worlds.

Incoherent agents with stable updating plans have at least one alternative, better

credal strategy with the same plans. This does not mean the the agent should change their initial credences to adopt a dominating credal strategy, but that this is possible, while keeping the same updating plans. Conversely, incoherent agents with unstable updating plans could achieve more accuracy in all possible worlds only if their plans are changed.

In order for an incoherent agent to satisfy $\mathfrak{I}$-stability, their credal strategy $\Gamma = \Gamma_{cr} \cup \Gamma_{plan}$ must be $\mathfrak{I}$-accuracy-dominated by a condionalizing $\Psi = \Psi_{cr} \cup \Gamma_{plan}$. But $\Psi_{cr} \cup \Gamma_{plan}$ $\mathfrak{I}$-accuracy-dominates $\Gamma_{cr} \cup \Gamma_{plan}$ iff[16] Γ_{cr} is dominated by Ψ_{cr}. Predd et al. (2009) show that, if we measure distance[17] between credence assignments with a divergence $d_{\mathfrak{I}}$ corresponding[18] to an additive inaccuracy measure $\mathfrak{I}$ based on a proper scoring rule, the coherent credences Γ_{cr}^* that are closest to the incoherent, unconditional credences Γ_{cr} will strongly $\mathfrak{I}$-accuracy-dominate it. Therefore, given incoherent credences Γ_{cr}, the agent can always form $\mathfrak{I}$-stable updating plans by conditionalizing this closest coherent Γ_{cr}^*. By Theorem 1 (see Appendix), these credal strategies are exactly those with null plan $\mathcal{I}_{d_{\mathfrak{I}}}$-incoherence, where $\mathcal{I}_{d_{\mathfrak{I}}}$ measures incoherence as the distance to the $d_{\mathfrak{I}}$-closest coherent credences. Thus, for unconditional initial credences, an incoherent agent satisfying $\mathcal{I}_{d_{\mathfrak{I}}}$-$TC$ is guaranteed to have $\mathfrak{I}$-stable plans.[19]

If the agent wants to satisfy a specific $\mathcal{I}_{d_{\mathfrak{I}}}$-$TC$ in order to obtain the corresponding epistemic benefit of $\mathfrak{I}$-stable updating plans, they might have to violate standard TC, being vulnerable to larger Dutch strategy losses. Particularly, TC is $\mathcal{I}_{d_1}$-TC for unconditional initial credences, but since there is no $\mathfrak{I}$ based on a proper scoring rule that yields $d_{\mathfrak{I}} = d_1$, the closest credences according to d_1 and $d_{\mathfrak{I}}$ in general differ. But fortunately, in some cases the agent can avoid this dilemma. For instance, if the agent assigns their initial credences to a partition, then for any credence function and any additive inaccuracy measure $\mathfrak{I}$ based on a proper scoring rule, it is possible to satisfy both $\mathcal{I}_{d_{\mathfrak{I}}}$-$TC$ and standard TC, where $d_{\mathfrak{I}}$ is the divergence corresponding to $\mathfrak{I}$ (see Proposition 2 in the Appendix). In other words, if $\mathfrak{I}$ is an additive inaccuracy measure based on a proper scoring rule, and if initial credences are assigned over a partition, there is a set of updating plans with zero plan $\mathcal{I}_{d_{\mathfrak{I}}}$-incoherence and zero plan incoherence. As in general, for any incoherent set of unconditional credences, there is a set of Manhattan-closest probability measures, and TC implies conditionalizing some measure in this set, this compatibility with an arbitrary $\mathcal{I}_{d_{\mathfrak{I}}}$-$TC$ might hold even in cases where credences are not defined over a partition.

---

[16]It is assumed that the inaccuracy measure $\mathfrak{I}$ satisfies Temporal Separability (Briggs & Pettigrew, 2020): $\mathfrak{I}(\Gamma, w) = \mathfrak{I}(\Gamma_{cr}, w) + \mathfrak{I}(\Gamma_{plans}, w)$ for any possible world w.

[17]Distance measures must satisfy symmetry and the triangle inequality, and some divergences $d_{\mathfrak{I}}$ do not, so we are abusing the terminology here.

[18]See for instance Gneiting and Raftery (2007) for the definition of this correspondence.

[19]For more details on formulating accuracy-based incoherence measures, see De Bona and Staffel (2017).

Example 10. *Recall Jane's credal strategy, from example 6, focusing on the initial cre-dences in the partition $A\&B, \neg A\&B, \neg B$:*

$$J_{cr} = \{cr(A\&B) = 0.4, cr(\neg A\&B) = 0.4, cr(\neg B) = 0.1\}, J_{plan} = \{cr_B(A) = 0.5\}$$

Let $\mathfrak{I}$ be the additive inaccuracy measure based on the Brier Score, so that the correspond-ing divergence $d_{\mathfrak{I}}$ is (squared) Euclidean distance. The coherent credences J_{cr}^* that are Euclidean-closest to J_{cr} are:

$$J_{cr}^* = \{cr(A\&B) = 0.433, cr(\neg A\&B) = 0.433, cr(\neg B) = 0.133\}$$

As J_{cr}^* implies $cr(A|B) = 0.5$, conditionalizing it yields $cr_B(A) = 0.5$. Thus, by Theorem 1, $J_{cr} \cup J_{plan}$ has null plan $\mathcal{I}_{d_{\mathfrak{I}}}$-incoherence, and Jane satisfies $\mathcal{I}_{d_{\mathfrak{I}}}$-TC. Furthermore, the Euclidean-closest coherent assignment, J_{cr}^* Brier-dominates J_{cr}, entailing that the credal strategy $J_{cr}^* \cup J_{plan}$ Brier-dominates $J_{cr} \cup J_{plan}$, which means that the plans in J_{plan} are Brier-stable. Finally, note that J_{cr}^* are also Manhattan-closest to J_{cr}, and, by Corollary 1, Jane thus satisfies standard TC.

7 Conclusion

Our aim in this paper was to answer the question of how agents with incoherent initial credences should update when they learn new evidence. We wanted to identify a desirable property of updates, such that this property would hold of conditionalizing updating plans for coherent agents, and of some privileged set of updating plans for incoherent agents. To characterize such a property, we started by choosing an incoherence measure that tracks normalized Dutch book loss, and showed how to define an agent's *plan incoherence*, which is the additional incoherence stemming from adding the agent's updating plans to their initial, possibly incoherent unconditional credences. In defining *plan incoherence*, we took advantage of the fact that a Dutch strategy involving updating plans always corresponds to a Dutch book involving conditional bets. Inspired by previous work on how to identify desirable properties of augmentative reasoning in incoherent agents, we then defined updating plans as satisfying *tolerant conditionalization* just in case their plan incoherence is zero. This property is satisfied only by conditionalizing updating plans when agents are initially coherent. When agents are initially incoherent, TC essentially asks them to follow updating plans that cohere with some closest coherent credence assignment identified by our incoherence measure. While we think TC identifies a desirable property of updates for incoherent agents, we don't think updating by TC has the status of a requirement. Sometimes there can be overriding reasons to choose an alternative updating plan, for example when an agent is aware of an epistemically superior way of fixing their

incoherence.

We then demonstrated some ways of extending and generalizing TC. We originally assumed that a credal strategy only contained an agent's initial unconditional credences and their updating plans. We showed how to incorporate an agent's conditional credences (which, in incoherent agents may or may not correspond to their updating plans). We then demonstrated that for incoherent agents, complying with TC sometimes requires not updating according to the ratio formula $cr_B(A) = cr(A\&B)/cr(B)$. Also, agents are sometimes better off (in terms of Dutch book loss) if they have conditional credences violating the ratio formula, independently of any updates. Lastly, we showed how to define alternative versions of TC by using different incoherence measures. For example, one might use a different type of Dutch book measure, or an incoherence measure that tracks accuracy dominance relations between credence functions. Different measures will often recommend different updating plans as satisfying the corresponding version of TC. Yet, we showed that when assigning credences to a partition, agents can always satisfy standard TC and some version of TC that tracks accuracy dominance simultaneously.[20]

8   Appendix

We say a probability measure $pr : \mathcal{L} \rightarrow [0,1]$ *satisfies* a credence $cr(A|B) = q$ if $pr(A\&B) = pr(B) \times q$. For unconditional credences, $B = \top$ and $pr(A\&B) = pr(B) \times q$ is equivalent to $pr(A) = q$. A probability measure satisfies a set of credences (or, equivalently, a credence assignment) Γ if it satisfies all credences in Γ.

The main technical results are stated for incoherence measures $\mathcal{I}_d$ defined via functions d that measure the distance from a given set of credences Γ to a probability measure pr:

$$\mathcal{I}_d(\Gamma) = \min\{d(\Gamma, pr)|pr \text{ is a probability measure}\}$$

The function d is assumed to be *adequate*, meaning it satisfies four properties:

- $d(\Gamma, pr) \geq 0$;

- $d(\Gamma, pr) = 0$ iff pr satisfies Γ;

- $d(\Gamma, pr) \leq d(\Gamma \cup \Psi, pr)$;

- $d(\Gamma \cup \Psi, pr) = d(\Gamma, pr)$ iff pr satisfies Ψ.

The first and the second properties ensure that d behaves like a distance, being zero iff the credences Γ agree with pr. The third one guarantees that the measure $\mathcal{I}_d$ is monotonic,

avoiding decreases in incoherence when we add credences. The fourth property allows the agent to assign new credences without increasing their incoherence by following the closest probability measure pr; and this is the only way to achieve that.

An inaccuracy measure $\mathfrak{I}$ is *additive* if $\mathfrak{I}(cr, w) = \sum_i S(cr(A_i), w)$ for a scoring rule S. For a coherent credence assignment cr', $\mathfrak{I}(cr, cr')$ denotes the expected value of $\mathfrak{I}(cr, w)$ according to cr'. For an inaccuracy measure $\mathfrak{I}$, the corresponding divergence is defined as $d_{\mathfrak{I}}(cr, cr') = \mathfrak{I}(cr, cr') - \mathfrak{I}(cr', cr')$.

We point out that, for any divergence D that measures the distance between (conditional) credence assignments over the same sentences, we can define a function d over credence assignments and probability measures as:

$$d(\Gamma, pr) = \min\{D(\Gamma, \Psi) | pr \text{ satisfies } \Psi\}$$

When $pr(B) = 0$, any $cr(A|B) = q$ is satisfied by pr, but the minimization guarantees that the conditional credences $cr(.|B) = q \in \Psi$ match those in Γ. Now the incoherence measure $\mathcal{I}_d$ is equivalent to the distance to the D-closest coherent credence assignment over the same sentences, as defined in the main text for simplicity:

$$\mathcal{I}_d(\Gamma) = \min\{D(\Gamma, \Psi) | \Psi \text{ is coherent}\}$$

Note that, if $D = d_{\mathfrak{I}}$ for an additive $\mathfrak{I}$ based on a proper scoring rule, the corresponding d is adequate.

Theorem 1. *A credal strategy $\Gamma = \Gamma_{cr} \cup \Gamma_{plan}$ yields null plan $\mathcal{I}_d$-incoherence, for an adequate d, iff there is a probability measure $pr : \mathcal{L} \to [0, 1]$ satisfying Γ_{plan}^{sync} such that $\mathcal{I}_d(\Gamma_{cr}) = d(\Gamma_{cr}, pr)$.*

Proof. ($\to$) To prove the contrapositive, suppose there is no probability measure pr satisfying Γ_{plan}^{sync} such that $\mathcal{I}_d(\Gamma_{cr}) = d(\Gamma_{cr}, pr)$. Then, for any probability measure pr, either $pr(A\&E) \neq q \times pr(E)$ for some $cr_E(A) = q \in \Gamma_{plan}$, and pr does not satisfy Γ_{plan}^{sync}, or $\mathcal{I}_d(\Gamma_{cr}) < d(\Gamma_{cr}, pr)$. In both cases, since d is adequate, we have $d(\Gamma^{sync}, pr) > \mathcal{I}_d(\Gamma_{cr})$. Therefore, $\mathcal{I}_d(\Gamma^{sync}) > \mathcal{I}_d(\Gamma_{cr})$, and plan $\mathcal{I}_d$-incoherence is positive.

($\leftarrow$) Now suppose there is a probability measure pr satisfying Γ_{plan}^{sync} such that $\mathcal{I}_d(\Gamma_{cr}) = d(\Gamma_{cr}, pr)$. Thus, since d is adequate, $d(\Gamma^{sync}, pr) = d(\Gamma_{cr}, pr) = \mathcal{I}_d(\Gamma_{cr})$. As $\mathcal{I}_d$ is monotonic, $\mathcal{I}_d(\Gamma_{cr}) \leq \mathcal{I}_d(\Gamma^{sync})$, and $\mathcal{I}_d(\Gamma^{sync}) \leq d(\Gamma^{sync}, pr)$ implies $\mathcal{I}_d(\Gamma^{sync}) = \mathcal{I}_d(\Gamma cr)$. □

Corollary 1. *An unconditional credal strategy $\Gamma = \Gamma_{cr} \cup \Gamma_{plan}$ satisfies tolerant conditionalization iff there is a probability measure pr satisfying Γ_{plan}^{sync} and extending a credence assignment cr^* that is Manhattan-closest to Γ_{cr}.*

Proof. Recall that, for unconditional credences, the Manhattan distance from Γ_{cr} to a probabilistic credence assignment cr' is the 1-norm of the corresponding violations

vector $\langle \varepsilon_1, \ldots, \varepsilon_n \rangle$, with $n = |\Gamma_{cr}|$, where the violations are computed via any probability measure extending cr'. Consider a function d such that $d(\Gamma, pr) = ||\langle \varepsilon_1, \ldots, \varepsilon_n \rangle||_1$, noting it is adequate. We have that $\mathcal{I}_\varepsilon(\Gamma_{cr}) = \mathcal{I}_d(\Gamma_{cr})$. For a probability measure pr, $\mathcal{I}_d(\Gamma_{cr}) = d(\Gamma_{cr}, pr)$ iff pr extends a credence assignment cr^* that is Manhattan-closest to Γ_{cr}. The result then follows from Theorem 1, as TC is equivalent to null plan $\mathcal{I}_\varepsilon$-incoherence. $\square$

Proposition 1. *Let $d_\mathfrak{I}$ be the divergence corresponding to an additive inaccuracy measure $\mathfrak{I}$ based on a proper scoring rule. Let a, b, c be coherent credence assignments over a set of sentences $\mathcal{A} \subseteq \mathcal{L}$, with $a \neq c$. If $b(A) = \lambda a(A) + (1 - \lambda)c(A)$ for all $A \in \mathcal{A}$ and some $\lambda \in (0, 1]$, then $d_\mathfrak{I}(a, b) < d_\mathfrak{I}(a, c)$.*

Proof. By definition, $d_\mathfrak{I}(a, b) = \mathfrak{I}(a, b) - \mathfrak{I}(b, b)$. Since $\mathfrak{I}(., b)$ is linear in b,

$$d_\mathfrak{I}(a, b) = \lambda \mathfrak{I}(a, a) + (1 - \lambda)\mathfrak{I}(a, c) - \lambda \mathfrak{I}(b, a) - (1 - \lambda)\mathfrak{I}(b, c)$$

For $\mathfrak{I}$ is proper, $-\mathfrak{I}(b, a) \leq -\mathfrak{I}(a, a)$ and $-\mathfrak{I}(b, c) < -\mathfrak{I}(c, c)$, implying:

$$d_\mathfrak{I}(a, b) < \lambda \mathfrak{I}(a, a) + (1 - \lambda)\mathfrak{I}(a, c) - \lambda \mathfrak{I}(a, a) - (1 - \lambda)\mathfrak{I}(c, c)$$

$$d_\mathfrak{I}(a, b) < (1 - \lambda)(\mathfrak{I}(a, c) - \mathfrak{I}(c, c)) = (1 - \lambda)d_\mathfrak{I}(a, c)$$

As $\lambda \in (0, 1]$ and $d_\mathfrak{I}(a, c) > 0$, we have $d_\mathfrak{I}(a, b) < d_\mathfrak{I}(a, c)$, finishing the proof. $\square$

Proposition 2. *Let $d_\mathfrak{I}$ be the divergence corresponding to an additive inaccuracy measure $\mathfrak{I}$ based on a proper scoring rule. If initial credences are unconditional and assigned to sentences that form a partition, it always possible to satisfy TC while having null plan $\mathcal{I}_{d_\mathfrak{I}}$-incoherence.*

Proof. Consider an unconditional credence assignment $cr : \mathcal{P} \to [0, 1]$ over a partition $\mathcal{P} = \{A_1, \ldots, A_m\}$. Let $cr_\mathfrak{I} : \mathcal{P} \to [0, 1]$ be such that $\mathcal{I}_{d_\mathfrak{I}}(cr) = d_\mathfrak{I}(cr, cr_\mathfrak{I})$. We will prove by contradiction that $\mathcal{I}_\varepsilon(cr) = d_1(cr, cr_\mathfrak{I})$, hence we suppose $\mathcal{I}_\varepsilon(cr) < d_1(cr, cr_\mathfrak{I})$. The result then follows by Theorem 1 and Corollary 1, for conditionalizing $cr_\mathfrak{I}$ would yield $\mathcal{I}_{d_\mathfrak{I}}$-$TC$ and TC.

If cr is coherent, conditionalizing entails TC and null plan $\mathcal{I}_{d_\mathfrak{I}}$-incoherence, so we focus on the two incoherent cases, where $\sum_i cr(A_i) < 1$ and $\sum_i cr(A_i) > 1$.

Consider first that $\sum_i cr(A_i) < 1$. Any credence assignment $cr^* : \mathcal{A} \to [0, 1]$ is a coherent Manhattan-closest one iff $cr^*(A_i) \geq cr(A_i)$ for $1 \leq i \leq m$ and $\sum_i cr^*(A_i) = 1$. As $\mathcal{I}_\varepsilon(cr) < d_1(cr, cr_\mathfrak{I})$, we have that $cr_\mathfrak{I}(A_i) < cr(A_i)$ for at least some i. Let $\mathcal{M}$ be the non-empty set of propositions A_i with $cr_\mathfrak{I}(A_i) \leq cr(A_i)$. Note that $\mathcal{M} \subsetneq \mathcal{P}$, for $\sum_i cr(A_i) < \sum_i cr_\mathfrak{I}(A_i)$ implies $cr_\mathfrak{I}(A_i) > cr(A_i)$ for some i. Let $cr_\mathfrak{I}^- : \mathcal{M} \to [0, 1]$ and $cr_\mathfrak{I}^+ : \mathcal{P} \setminus \mathcal{M} \to [0, 1]$ be credence assignments such that $cr_\mathfrak{I}^-(A_i) = cr_\mathfrak{I}(A_i)$ for all $A_i \in \mathcal{M}$ and $cr_\mathfrak{I}^+(A_i) = cr_\mathfrak{I}(A_i)$ for all $A_i \in \mathcal{P} \setminus \mathcal{M}$. Analogously, consider $cr^- : \mathcal{M} \to [0, 1]$ and

$cr^+ : \mathcal{P} \setminus \mathcal{M} \to [0,1]$, just narrowing the domain of cr. As $\sum_{\mathcal{M}} cr(A_i) + \sum_{\mathcal{P} \setminus \mathcal{M}} cr_{\mathfrak{I}}(A_i) > 1 > \sum_{\mathcal{P}} cr(A_i)$, there is a $\lambda \in (0,1)$ such that $\sum_{\mathcal{P} \setminus \mathcal{M}} \lambda cr(A_i) + (1 - \lambda) cr_{\mathfrak{I}}(A_i) = 1 - \sum_{\mathcal{M}} cr(A_i)$. Define $c_\lambda^+ : \mathcal{P} \setminus \mathcal{M} \to [0,1]$ via $cr_\lambda^+(A_i) = \lambda cr(A_i) + (1 - \lambda) cr_{\mathfrak{I}}(A_i)$ for all $A_i \in \mathcal{P} \setminus \mathcal{M}$. Let $c_\lambda : \mathcal{P} \to [0,1]$ be the union between cr^- and cr_λ^+. Note that c_λ, c_λ^+, c_λ^-, $cr_{\mathfrak{I}}$ and $cr_{\mathfrak{I}}^+$ are coherent. As $\mathfrak{I}$ is additive, $d_{\mathfrak{I}}(cr, cr_\lambda) = d_{\mathfrak{I}}(cr^-, cr_\lambda^-) + d_{\mathfrak{I}}(cr^+, cr_\lambda^+) = d_{\mathfrak{I}}(cr^+, cr_\lambda^+)$. By Proposition 1, $d_{\mathfrak{I}}(cr^+, cr_\lambda^+) < d_{\mathfrak{I}}(cr^+, cr_{\mathfrak{I}}^+) \le d_{\mathfrak{I}}(cr, cr_{\mathfrak{I}})$. Hence, $cr_{\mathfrak{I}}$ is not the $d_{\mathfrak{I}}$-closest coherent credence assignment, which is a contradiction.

If $\sum_i cr(A_i) > 1$, the proof is completely symmetrical. Now, $\mathcal{I}_\varepsilon(cr) < d_1(cr, cr_{\mathfrak{I}})$ implies $cr(A_i) < cr_{\mathfrak{I}}(A_i)$ for some A_i and $cr(A_i) > cr_{\mathfrak{I}}(A_i)$ for other A_j. We define in the same way the set $\mathcal{M}$ and the credence assignments $cr_{\mathfrak{I}}^+$, $cr_{\mathfrak{I}}^-$, cr^+ and cr^-. Now cr_λ^- is a convex combination between $cr_{\mathfrak{I}}^-$ and cr^-, such that its union with cr^+ forms the coherent c_λ. As $\mathfrak{I}$ is additive, $d_{\mathfrak{I}}(cr, cr_\lambda) = d_{\mathfrak{I}}(cr^-, cr_\lambda^-)$. By Proposition 1, $d_{\mathfrak{I}}(cr^-, cr_\lambda^-) < d_{\mathfrak{I}}(cr^-, cr_{\mathfrak{I}}^-) \le d_{\mathfrak{I}}(cr, cr_{\mathfrak{I}})$, which is a contradiction, finishing the proof. $\square$

Theorem 2. *Let $\Gamma_{cr} = \{cr(A_j) = q_j | 1 \le j \le m\}$ be a set of unconditional credences, where $\{A_1, \ldots, A_m\} \subseteq \mathcal{L}$ can be partitioned into sets of sentences forming partitions. Let $\Pi_\ge$ (respectively, $\Pi_\le$) be the set of probability measures $pr : \mathcal{L} \to [0,1]$ with $pr(A_j) \ge q_j$ ($pr(A_j) \le q_j$) for all $1 \le j \le m$. When $\Pi_\ge$ ($\Pi_\le$) is non-empty, a credal strategy $\Gamma_{cr} \cup \Gamma_{plan}$ satisfies TC iff there is a $pr \in \Pi_\ge$ ($pr \in \Pi_\le$) satisfying Γ_{plan}^{sync}.*

Proof. Assume $\Pi_\ge$ (respectively, $\Pi_\le$) is non-empty. By Corollary 1, it suffices to prove that $pr \in \Pi_\ge$ ($\Pi_\le$) iff pr extends a coherent credence assignment that is Manhattan-closest to Γ_{cr}. We start by partitioning $\mathcal{A} = \{A_1, \ldots, A_m\}$ into $\bigcup_{i=1}^{n} \mathcal{B}_i$, where each $\mathcal{B}_i \subseteq \mathcal{A}$ is a set of sentences forming a partition. Let $\Gamma_i \subseteq \Gamma_{cr}$ be $\{cr(A_j) = q_j | A_j \in \mathcal{B}_i\}$, such that $\Gamma_{cr} = \Gamma_1 \cup \cdots \cup \Gamma_n$. For each $1 \le i \le n$ and any probability measure pr, we have $\sum_{A_j \in \mathcal{B}_i} pr(A_j) = 1$, since $\mathcal{B}_i$ forms a partition. As $\Pi_\ge$ ($\Pi_\le$) is non-empty, it follows, for all i, that $\sum_{A_j \in \mathcal{B}_i} q_j \le 1$ (≥ 1). Consequently, $\mathcal{I}_\varepsilon(\Gamma_i) = 1 - \sum_{A_j \in \mathcal{B}_i} q_j$ (respectively, $= -1 + \sum_{A_j \in \mathcal{B}_i} q_j$) for each $1 \le i \le n$. Note that, in general, we must have that $\mathcal{I}_\varepsilon(\Gamma_{cr}) \ge \sum_{i=1}^{n} \mathcal{I}_\varepsilon(\Gamma_i)$.

Consider a probability measure $pr \in \Pi_\ge$ ($\Pi_\le$). Note that $\sum_{A_j \in \mathcal{B}_i} |pr(A_j) - q_j| = \sum_{A_j \in \mathcal{B}_i} pr(A_j) - q_j$ (resp., $= \sum_{A_j \in \mathcal{B}_i} q_j - pr(A_j)$), thus $\sum_{A_j \in \mathcal{B}_i} |pr(A_j) - q_j| = \mathcal{I}_\varepsilon(\Gamma_i)$. Consequently, as $\mathcal{I}_\varepsilon(\Gamma_{cr}) \le \sum_{i=1}^{n} \sum_{A_j \in \mathcal{B}_i} |pr(A_j) - q_j| = \sum_{i=1}^{n} \mathcal{I}_\varepsilon(\Gamma_i)$, we have that $\mathcal{I}_\varepsilon(\Gamma_{cr}) = \sum_{i=1}^{n} \mathcal{I}_\varepsilon(\Gamma_i)$. Therefore, $\mathcal{I}_\varepsilon(\Gamma_{cr}) = \sum_{j=1}^{m} |pr(A_j) - q_j|$ and pr extends a coherent credence assignment over $\mathcal{A}$ that is Manhattan-closest to Γ_{cr}.

Consider now a probability measure $pr : \mathcal{L} \to [0,1]$ with $pr(A_l) < q_l$ (respectively, $pr(A_l) > q_l$) for some $cr(A_l) = q_l \in \Gamma_k$, for a $\Gamma_k \subseteq \Gamma_{cr}$. This implies $\sum_{A_j \in \mathcal{B}_k} pr(A_j) - q_j <$

$\sum\limits_{A_j \in \mathcal{B}_k} |pr(A_j) - q_j|$ (resp. $\sum\limits_{A_j \in \mathcal{B}_k} q_j - pr(A_j) < \sum\limits_{A_j \in \mathcal{B}_k} |q_j - pr(A_j)|$), which means $\mathcal{I}_\varepsilon(\Gamma_k) <$ $\sum\limits_{A_j \in \mathcal{B}_k} |pr(A_j) - q_j|$. Hence, $\mathcal{I}_\varepsilon(\Gamma_{cr}) = \sum\limits_{i=1}^{n} \mathcal{I}_\varepsilon(\Gamma_i)$ implies $\mathcal{I}_\varepsilon(\Gamma_{cr}) < \sum\limits_{j=1}^{m} |pr(A_j) - q_j|$. It follows that pr does not extend a Manhattan-closest coherent credence assignment. $\square$

Corollary 2. *Let $\Gamma_{cr} = \{cr(A_j) = q_j | 1 \leq j \leq m\}$ be a set of unconditional credences, where $\{A_1, \ldots, A_m\} \in \mathcal{L}$ forms an algebra. Let $\Pi_\geq$ (respectively, $\Pi_\leq$) be the set of probability measures $pr : \mathcal{L} \to [0,1]$ with $pr(A_j) \geq q_j$ ($pr(A_j) \leq q_j$) for all $1 \leq j \leq m$. When $\Pi_\geq$ ($\Pi_\leq$) is non-empty, a credal strategy $\Gamma_{cr} \cup \Gamma_{plan}$ satisfies TC iff there is a $pr \in \Pi_\geq$ ($pr \in \Pi_\leq$) satisfying Γ_{plan}^{sync}.*

Proof. As $\mathcal{A} = \{A_1, \ldots, A_m\}$ forms an algebra, we can partition it into pairs of sentences $\mathcal{B}_i = \{B_i, B_i'\}$, for $1 \leq i \leq m/2 = n$, such that B_i' is equivalent to $\neg B_i$, yielding $\mathcal{A} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_n$. The result then follows from Theorem 2. $\square$

References

Babic, B. (2020). Approximate coherentism and luck. *Philosophy of Science*, *forthcoming*, 1 – 38. doi: 10.1086/713905

Briggs, R. (2009). Distorted reflection. *Philosophical Review*, *118*(1), 59–85. doi: 10.1215/00318108-2008-029

Briggs, R. A., & Pettigrew, R. (2020). An accuracy-dominance argument for conditionalization. *Noûs*, *54(1)*, 162–181. doi: 10.1111/nous.12258

Christensen, D. (2004). *Putting Logic in its Place: Formal Constraints on Rational Belief*. Oxford University Press.

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106 – 133. doi: 10.1016/j.cogpsych.2016.06.006

Crupi, V., Fitelson, B., & Tentori, K. (2007). Probability, confirmation, and the conjunction fallacy. *Thinking and Reasoning*, *14*(2), 182–199. doi: 10.1080/13546780701643406

Das, N. (2020). Externalism and exploitability. *Philosophy and Phenomenological Research*. online first. doi: 10.1111/phpr.12742

De Bona, G., & Finger, M. (2015). Measuring inconsistency in probabilistic logic: rationality postulates and Dutch book interpretation. *Artificial Intelligence*, *227*, 140–164. doi: 10.1016/j.artint.2015.06.006

De Bona, G., & Staffel, J. (2017). Graded incoherence for accuracy-firsters. *Philosophy of Science*, *84*(2), 189–213. doi: 10.1086/690715

De Bona, G., & Staffel, J. (2018). Why be (approximately) coherent? *Analysis*, *78*(3), 405–415. doi: 10.1093/analys/anx159

Diaconis, P., & Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association*, *77*(380), 822–830. doi: 10.1080/01621459.1982.10477893

Gallow, J. D. (2019a). Diachronic dutch books and evidential import. *Philosophy and Phenomenological Research*, *99*(1), 49–80. doi: 10.1111/phpr.12471

Gallow, J. D. (2019b). Updating for externalists. *Noûs*. online first. doi: 10.1111/nous.12307

Gilboa, I., & Schmeidler, D. (1993). Updating ambiguous beliefs. *Journal of Economic Theory*, *59*(1), 33–49. doi: 10.1006/jeth.1993.1003

Gilboa, I., & Schmeidler, D. (1994). Additive representations of non-additive measures and the choquet integral. *Annals of Operations Research*, *52*(1), 43–65. doi: 10.1007/BF02032160

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. doi: 10.1198/016214506000001437

Greaves, H., & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, *115*(459), 607–632. doi: 10.1093/mind/fzl607

Greco, D. (2017). Cognitive mobile homes. *Mind*, *126*(501), 93–121. doi: 10.1093/mind/fzv190

Hild, M. (1998). The coherence argument against conditionalization. *Synthese*, *115*(2), 229–258. doi: 10.1023/A:1005082908147

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. doi: 10.1037/h0034747

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–17. doi: 10.1017/s0140525x00041157

Nielsen, M. (2021). Accuracy-dominance and conditionalization. *Philosophical Studies*. online first. doi: 10.1007/s11098-020-01598-6

Oddie, G. (1997). Conditionalization, cogency, and cognitive value. *British Journal for the Philosophy of Science*, *48*(4), 533–541. doi: 10.1093/bjps/48.4.533

Pettigrew, R. (2020). *Dutch Book Arguments*. Cambridge, UK: Cambridge University Press. doi: 10.1017/9781108581813

Potyka, N. (2014). Linear programs for measuring inconsistency in probabilistic logics. In *Fourteenth International Conference on Principles of Knowledge Representation and Reasoning (KR-14)*. AAAI.

Predd, J. B., Seiringer, R., Lieb, E. H., Osherson, D. N., Poor, H. V., & Kulkarni, S. R. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, *55*(10), 4786–4792. doi: 10.1109/TIT.2009.2027573

Rescorla, M. (2020). An improved dutch book theorem for conditionalization. *Erkenntnis.*, (pp. 1–29). online first. doi: 10.1007/s10670-020-00228-1

Salow, B. (2019). Elusive externalism. *Mind*, *128*(510), 397–427. doi: 10.1093/mind/fzx015

Schervish, M. J., Kadane, J. B., & Seidenfeld, T. (2003). Measures of incoherence: How not to gamble if you must. In *Bayesian Statistics 7: Proceedings of the 7th Valencia Conference on Bayesian Statistics*, (pp. 385–402).

Schervish, M. J., Seidenfeld, T., & Kadane, J. B. (2002). Measuring incoherence. *Sankhyā: The Indian Journal of Statistics, Series A*, (pp. 561–587).

Schoenfield, M. (2017). Conditionalization does not (in general) maximize expected accuracy. *Mind*, *126*(504), 1155–1187. doi: 10.1093/mind/fzw027

Shafer, G. (1985). Conditional probability. *International Statistical Review / Revue Internationale de Statistique*, *53*(3), 261–275. doi: 10.2307/1402890

Staffel, J. (2019). *Unsettled Thoughts: A Theory of Degrees of Rationality*. Oxford University Press. doi: 10.1093/oso/9780198833710.001.0001

Teller, P. (1973). Conditionalization and observation. *Synthese*, *26*(2), 218–258. doi: 10.1007/bf00873264

Thimm, M. (2013). Inconsistency measures for probabilistic logics. *Artificial Intelligence*, *197*, 1 – 24. doi: 10.1016/j.artint.2013.02.001

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. doi: 10.1037/0033-295X.90.4.293

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567. doi: 10.1037/0033-295x.101.4.547

Van Boven, L., & Epley, N. (2003). The unpacking effect in evaluative judgments: When the whole is less than the sum of its parts. *Journal of Experimental Social Psychology*, *39*(3), 263 – 269. doi: 10.1016/S0022-1031(02)00516-4

Wang, G., Kulkarni, S. R., Poor, H. V., & Osherson, D. N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, *8*(2), 128–144. doi: 10.1287/deca.1110.0206

Zendejas Medina, P. (2021). Just as planned: Bayesianism, externalism, and plan coherence. *Manuscript*.

Zhao, J., Crupi, V., Tentori, K., Fitelson, B., & Osherson, D. (2012). Updating: Learning versus supposing. *Cognition*, *124*(3), 373–378. doi: 10.1016/j.cognition.2012.05.001

Zhao, J., Shah, A., & Osherson, D. (2009). On the provenance of judgments of conditional probability. *Cognition*, *113*(1), 26–36. doi: 10.1016/j.cognition.2009.07.006