Reflection, Introspection, and Book

Kevin Zollman Kevin Dorst*

December 5, 2024

Abstract

The much-debated Reflection principle states that a coherent agent's credences must match their estimates for their future credences. Defenders claim that there are Dutch-book arguments in its favor, putting it on the same normative footing as probabilistic coherence. Critics claim that those arguments rely on the implicit, implausible assumption that the agent is *introspective*: that they are certain what their own credences are. In this paper, we clarify this debate by surveying several different conceptions of the book scenario. We show that the crucial disagreement hinges on whether agents who are not introspective are known to reliably act on their credences: if they *are*, then coherent Reflection failures are (at best) ephemeral; if they *aren't*, then Reflection failures can be robust—and perhaps rational and coherent. We argue that the crucial question for future debates is which notion of coherence makes sense for such unreliable agents, and sketch a few avenues to explore.

One version of the 'Reflection' principle in epistemology requires that an agent's current probability in some event E must be the expectation of her future credence in E. For instance, suppose her credence in some other event, F, is $C(F) = 0.5$. Suppose further that she expects that if she learns F, her future credence will be $C^+_F(E) = 0.6$, while if she learns $\neg F$, her credence will be $C^+_{\neg F}(E) = 0.2$. Reflection requires that her credence today must be a probability-weighted average of these two possibilities, which in this case is 0.4 $(= 0.5(0.6) + 0.5(0.2))$.

Reflection has been subject to much discussion.[1] Under standard Bayesian updating, it always holds (Kadane et al., 1996; Briggs, 2009). But several authors have explored types of probabilistic updating where it fails, and some argue that an agent who violates Reflection might nevertheless be coherent and rational.[2] There are many different concepts of coherence and rationality. In this paper we'll explore one: coherence as being free from (Dutch) book.

*Authors contributed equally, and are listed in reverse-alphabetical order.

[1] Classics: Skyrms 1980, 1990, 2006; van Fraassen 1984; Gaifman 1988; Christensen 1991.

[2] Cf. Williamson 2000, 2008; Elga 2013; Salow 2018; Gallow 2019, 2021; Dorst 2020a; Dorst et al. 2021; Das 2022; Dorst 2023; Rescorla 2023; van Fraassen 2023; Isaacs and Levinstein 2024.

We consider two agents who violate Reflection in similar ways, but for importantly different reasons. One does so because her opinions are *unstable*, while the other does so because she is not *introspective*—she is not aware of her own credences.

We consider three variations of the book argument: a version where the agent posts fair prices; one where the bookie must offer bets but can form conditional strategies in response to the agent's actions; and one where the bookie must commit to a set of offered-bets-and-prices beforehand.

The unstable agent—who violates Reflection while obeying introspection—is subject to book in all three settings. We conclude this agent is incoherent and (in many senses) irrational.

Things are more complicated for the non-introspective agent. Assuming that she is known to reliably act on her credences, in the first two book arguments, she faces a dilemma: either her violations of Reflection are ephemeral, or they are bookable. This dilemma isn't present for the third version of the book argument. And in all cases, the knowledge that she will reliably act on her credences plays a crucial role, for it implies that her actions reveal information about her beliefs. As we'll see, when this assumption holds there is strong pressure toward Reflection.

What to make of this? We suggest that under certain kinds of cognitive bounds, Reflection-violating agents might nonetheless be rational—but how principled those bounds are is up for debate. Although introspection-failures appear to be well-motivated, a recurrent theme of our discussion is that they can't stand on their own. In order for introspection failures to lead to *persistent* Reflection failures, they must be accompanied with some (believed) unreliability in the link between credences and action.

Work in non-introspective epistemology often discusses the idea that all levels of our cognitive systems are noisy and subject to error (Williamson, 2000), but this idea is rarely formalized or modeled explicitly. This explains the current dialectical impasse: defenders and critics of introspection and Reflection may be implicitly making different assumptions about what sorts of unreliability our credences and actions are subject to. We conclude by discussing a prominent way of modeling such unreliability—the 'sampling model' from cognitive science (Icard, 2016)—suggesting that it gives an example of how the link between credences and action might be (present but) unreliable, and thus introspection failures might be persistent. Although this idea is empirically plausible and holds some theoretical promise, it's difficult to formulate a notion of book and coherence for it. We'll conclude by sketching a few possibilities.

Ultimately, whether or not a violation of Reflection should be viewed as incoherent or irrational turns on several thorny philosophical debates that go beyond a simple analysis. One cannot easily conclude that such agents are coherent or incoherent, rational or irrational. It depends on *exactly* how and why they violate Reflection, what their abilities and bounds are, and what the right notion of book is for agents with those bounds. We don't claim that we've gotten to the bottom of this—but we do hope to clarify the issues and push the debate forward.

1 Violations of Reflection

Suppose there's a preexisting algebra of events $\mathcal{G}$ under discussion. The agent has some function $C : \mathcal{G} \to \mathbb{R}$ that represents her credences for events in that algebra—a quantitative measurement of the agent's degrees of belief. We don't yet call it a probability function, or impose any constraints on it, because it is exactly those constraints that are under discussion.

We'll consider situations where the agent is unsure of her own beliefs, so $\mathcal{G}$ must include events of the form "my credence in E is currently 0.5", "conditional on learning F, my credence in E tomorrow will be 0.6" and the like. This is done by specifying maximally-specific events as worlds that are specific enough to settle both (a) the truth values of a set of "extra agent" propositions ("the coin lands heads", etc.) and (b) what the agent's credences are at that world. To keep the discussion accessible, we'll largely suppress the formal details.[3]

Because we are interested in how an agent responds to information, we will need to talk about about the agent's credences "today" which we will denote by C and their credences "tomorrow," which we will denote by C^+. If an agent is expected to learn an event between today and tomorrow, we will also talk about C^+_E which specifies what the agent's credences will be after learning E. C^+_E says what her future credence function is if *in fact* E is true—we are not going to assume that the agent always updates by conditioning on E. (For example, if E occurs but has no influence on the agent whatsoever, C^+_E will just be C.) Since the agent doesn't know what evidence they'll receive, their prior C will be unsure whether there future credence function C^+ is equal to C^+_E or instead equal to $C^+_{\neg E}$.

The principle of Reflection can be stated as a relationship between one's current credence and one's future credence: one's current credence is the expectation of one's future credence. Here "expectation" is defined as the probability-weighted average, according to the prior C. Consider an agent who assigns fair odds on the flip of a coin: $C(H) = C(T) = 0.5$. Suppose that she knows she will learn the outcome of the coin and will update by conditioning on this fact. So, she expects that $C^+_H(H) = 1$ and $C^+_T(H) = 0$. Since she thinks each is equally likely, her expectation for her future credence in H (and T) equals 0.5—she obeys Reflection.

What about an agent who doesn't? Suppose our agent knows that tomorrow, if the coin comes up heads, she'll learn that fact and update. So in world H, she will assign $C^+_H(H) = 1$ and $C^+_H(T) = 0$. However, if the coin comes up tails she will fail to update. Her credences will remain the same, $C^+_T(H) = C^+_T(T) = 0.5$. Notice that the agent currently thinks that the probability that $C^+ = C^+_H$ is 0.5. The agent violates the Reflection principle because $C(H)$ does not equal her current expectation of her future credence in heads: $\mathbb{E}_C(C^+(H)) = C(H) \cdot 1 + C(T) \cdot 0.5 = 0.75$, which is greater than her current credence of $C(H) = 0.5$.

How might an agent come to occupy this position? Start with a quite pathological case. She might know all these things in advance and use them to con-

[3] But see e.g. Harsanyi 1967; Gärdenfors 1975; Gaifman 1988; Samet 2000; Williamson 2000, 2008 for applications of such models. Williamson 2019 and Dorst 2020b give summaries.

struct conditional probabilities. So, for this agent $C(H|T) = C^+_T(H) = 0.5$ and $C(T|H) = C^+_H(T) = 0$. She also knows that she will update by conditioning on one of H or T. Assuming that the agent obeys the "ratio formula" that $C(H|T) = C(H \cap T)/C(T)$, the agent cannot have a single value for $C(H \cap T)$.[4] While, of course, an agent might do this, we will ignore this possibility for future discussion because we will assume a minimal coherence condition, that $C(q)$ always takes a single value for given q, and that our agent's conditional credences obey the ratio formula.

More interestingly, the agent might construct synchronically coherent credences, but fail to update by conditioning on the true answer to the question, "How did the coin land?", i.e. the true member of {H, T}. This agent might assign $C(H|T) = 0$ (as she should), but recognize that tomorrow $C^+_T(H) = 0.5$. The question is what the agent models herself as knowing tomorrow about the situation—in particular what she knows about her own credences.

One possibility is that in the future she'll violate introspection. Formally, introspection is the axiom that if $C^+_i(q) = t$, then $C^+_i(C^+(q) = t) = 1$ (where, in our case, $i \in \{H, T\}$). Suppose the agent knows throughout that when the coin lands heads, she is certain of this (i.e. that $C^+_H(H) = 1$), and that when the coin lands tails, she's uncertain (i.e. that $C^+_T(H) = 0.5$). In order to know that throughout, we must construct a single C^+ which has such credences at each world. We can diagram her future credences as follows:

$$\begin{tikzcd} \mathrm{T} \arrow[loop above, "0.5"] \arrow[r, "0.5"] & \mathrm{H} \arrow[loop above, "1"] \end{tikzcd}$$

H is the possibility where the coin lands heads; T where it lands tails. The labeled arrows represent what, in that world, the agent's credences assign to each world: the fact that T has two '0.5' labeled arrows coming from it and going to each world says that $C^+_T(T) = 0.5$ and $C^+_T(H) = 0.5$.

In this model, the agent's posterior is not introspective. If the coin lands heads, she is certain that it landed heads: at H, $C^+(H) = 1$. When the coin lands tails, she is uncertain between being at the world (T) where it lands tails and at the world (H) where it lands heads *and where she knows it lands heads.* So although at T she *in fact* assigns 0.5 credence to H, she isn't certain of that fact: she's 50%-confident that she's *certain* of H.[5] Why can't she infer from the fact that she's uncertain between T and H to the conclusion that she's at T? Because she *isn't sure* that she's uncertain between T and H—as we just said, she's 50%-confident that instead she's certain she's at H. This is crucial: the only way to assign synchronically coherent credences to this situation—wherein she knows throughout that if H, she's certain of H, while if T, she's uncertain

[4] Why do we assume the agent obeys this definition? Because in our algebra there is no event "$C|T$", so we must relate this to some bet the agent might take. We could introduce "called-off" bets which allow for bets on conditionals, but we won't.

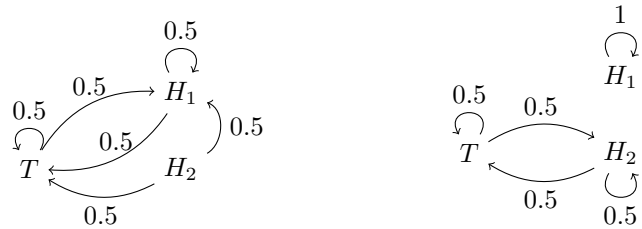
[5] Formally, the event $\langle C^+(H) = 1\rangle$ is the set of worlds where it's true—namely, $\{H\}$—while the event $\langle C^+(T) = 0.5\rangle$ is $\{T\}$. Thus $C^+_T(\langle C^+(H) = 1\rangle) = C^+_T(\{H\}) = 0.5$, while $C^+_T(\langle C^+(H) = 0.5\rangle) = C^+_T(\{T\}) = 0.5$.

between H and T—requires that the agent not know her own credences. We will refer to this violator of Reflection as the *non-introspective agent*.[6]

How would we model a similar update for an agent who's synchronically coherent *and always introspective*? Although she can't maintain all the knowledge about her dispositions throughout—for then she could figure out how the coin landed by seeing what her credences were—she can still know beforehand that 'If heads, I'll be sure of it; if tails, I'll be uncertain'. In order for this to happen, she needs to alter her view of the situation after updating.

To model this case properly—and keep track of her higher-order beliefs about her own beliefs—we need three possibilities, T, H_1, and H_2. T is the world where the coin lands tails but the agent continues to assign credence 0.5 to heads and 0.5 to tails. H_1 is where it lands heads and the agent becomes certain of heads: $C^+_{H_1}(H) = 1$. H_2 is a world where the coin lands heads, but the agent remains uncertain about whether the coin landed heads. Since our agent violates Reflection, her prior is credence is $C(T) = C(H_1) = 0.5$ and $C(H_2) = 0$: she's initially certain that she will not be in world H_2, but because she regards T and H_1 as possibilities, she doesn't know what the flip of the coin will be. Writing this as a 3-place vector over worlds (T, H_1, H_2), at each world w her prior is $C_w = (0.5, 0.5, 0)$. Meanwhile, her posterior, on the right of the below diagram, differs in different possibilities:

$$C = \left(\begin{array}{c|ccc} & T & H_1 & H_2 \\ \hline T & 0.5 & 0.5 & 0 \\ H_1 & 0.5 & 0.5 & 0 \\ H_2 & 0.5 & 0.5 & 0 \end{array}\right) \qquad C^+ = \left(\begin{array}{c|ccc} & T & H_1 & H_2 \\ \hline T & 0.5 & 0 & 0.5 \\ H_1 & 0 & 1 & 0 \\ H_2 & 0.5 & 0 & 0.5 \end{array}\right)$$



To read these diagrams, start with the matrices. The first column and row are labels. The rows say which probability function (over (T, H_1, H_2)) our agent

[6]Notice that a synchronic version of Reflection also fails in this example. At world T, conditional on her posterior assigning 0.5 to T, her posterior is certain that T is true: $C^+_T(T|C^+(T) = 0.5) = 1$. Why? Because she's sure that if T were false, she'd assign 0 to it: at all the H-worlds, $C^+(H) = 1$. In other words: at T she's 0.5 in T *only because* she's not sure that she's 0.5 in T—only because she's not introspective. This fact—that introspection failures lead to synchronic failures of Reflection—is what drives this type of non-introspective diachronic Reflection failure. Interestingly, the example generalizes completely: if you satisfy synchronic Reflection *for all propositions* at a world, then you must satisfy introspection at that world (see e.g. Samet 2000; Williamson 2000, 2008; Elga 2013; Lasonen-Aarnio 2014; Dorst 2020a). This fact is sometimes overlooked (e.g. Rescorla, 2023; van Fraassen, 2023).

has in each world, at each time. So the fact that every row in the C matrix is $(0.5, 0.5, 0)$ means that in each world she assigns 0.5 to each of T and H_1. Meanwhile, the fact that the first and third rows of C^+ are $(0.5, 0, 0.5)$ means that at both T and H_2, her posterior assigns 0.5 to T and H_2, and rules out H_1. This same information is captured in the 'Markov' (labeled-arrow) diagrams below, using the conventions from above.

This means that today the agent is certain that if the coin comes up heads, she'll shift to believing she is in H_1 with probability 1. Meanwhile, if the coin comes up tails, she will assign a chance to being in H_2, an event which she today thinks is impossible. Since she has the same distribution in worlds T and H_2, her credences are introspective: after the update, if T she knows she assigns $C(T) = 0.5$. We'll call this agent the *unstable agent* because she initially was certain that H_2 wouldn't happen, but is disposed to lose that certainty.

In the three sections that follow, we will consider how these two agents (unstable and non-introspective) fare with respect to three versions of the book story. We will find that the unstable agent can be subject to book in all three, demonstrating that this type of Reflection failure is incoherent and irrational.

The issue will be more complicated with the non-introspective agent. Given the first two setups of the book argument, there's a dilemma: either the agent will self-correct for her Reflection failures (making them transient), or she can be subject to these two versions of the book. The non-introspective agent is, however, immune from the third version ("fixed option") of the book argument. But as we'll see, in all cases the story assumes that the agent is known to reliably act on her credences—in §6 we'll come back to how this assumption interacts with our verdicts.

2 The sportsbook argument

We will start with the "sportsbook" version of the book argument.[7] The agent is the subject of our inquiry, the person who will be judged as coherent or incoherent. The other character is "the bookie," who is a fiction used for the purpose of illustration. The bookie represents decisions that the agent might face—a kind of worst-case scenario, not a prediction about what will really happen.

In the sportsbook version of the book argument, the agent acts like a casino sportsbook. For every event $E \in \mathcal{G}$, there are tickets that force the seller to pay the buyer \$1 if E occurs and are worthless if E does not. The agent posts prices for every event E on a board. Unlike a real casino, the bookie can either buy the tickets from the agent or sell them to the agent at the posted price.

The bookie observes the prices that the agent posts and then decides which tickets to buy and sell. If the bookie has a strategy that either strictly or weakly guarantees him a win, then the agent is said to be *incoherent*.[8] Don't take this

[7] Uchii 1973 gives an argument for introspection using a variant of the sportsbook setup.

[8] For our purposes we will only be dealing in strict books, but some want to include weak ones (where the bookie might win and cannot lose).

as necessarily a synonym for "irrational"—we'll return to this in section 5.

There are two constraints that are often added to this story. The first one is that the agent always posts her credences as prices: the price the agents posts for some event E is just $C(E)$. This is meant to prevent the agent from avoiding book by some mathematical sleight of hand.

Suppose, for example, the agent had incoherent credences $C(H) = 0.75$ and $C(T) = 0.75$ for the flip of a coin. If this agent posted these credences as prices, she would be easily booked: the bookie would sell one ticket on each of the two possible events for the price of \$0.75. The bookie would have to pay off only one of those tickets, netting a guaranteed profit of \$0.50.

One might try to save the agent by giving her a "price-posting strategy" where she posts only normalized prices. So her price for H is given by:

$$\frac{C(H)}{C(H) + C(T)} = 0.5$$

Of course, that will avoid book.

If she only behaves this way when in the book situation, then this is simply an attempt to dodge the book argument without really fixing the problem it was designed to diagnose. If, on the other hand, she does this *every* time she needs to engage in any action, then C is not really her credences—they are some other mathematical construct that she uses to generate her credences. So we assume that the agent's price-posting strategy is simply to post her credences, and that both the bookie and the agent know this.

The second constraint requires a little more discussion. The usual condition is that the bookie should not know more than the agent. This condition is imposed to prevent the agent from being judged as irrational unfairly. For example, if the bookie knows whether the coin landed heads—but the agent does not—then the bookie can buy tickets on heads with no risk of losing money. The bookie will book the agent, but only because he knows something the agent doesn't.

Although the constraint is usually described as requiring they have the same knowledge, this way of describing this constraint is both too restrictive and not restrictive enough. It's too restrictive because it pays attention to what the agent knows, not what she is a position to know. Suppose that the agent has the coin in front of her, has performed the flip, and is staring at the heads side of the coin—but for some reason, she hasn't updated her credence in heads. It seems strange to prevent the bookie from seeing the coin. The agent is being irrational by failing to update her credence in the coin—she has evidence which she's refused to use. So rather than saying the agent and the bookie know the same things, we should (at least) say that the bookie and the agent have the same evidence.

When we consider bounded agents, the "same evidence" condition should be further modified. Suppose we are considering agents who are bounded in that they cannot solve NP hard problems. It wouldn't be probative to note that they can be booked by a bookie who is capable of solving NP hard problems.

If we wanted to apply the book argument, we should also prevent the bookie from solving NP hard problems. The same motivation that made us add the "same evidence" criterion should extend to include any relevant form of cognitive bounds that we don't want to rule out as irrational. Another way to put this: for the book to result in a "sure loss", it needs to be knowable *from the agent's perspective* that it'll result in a loss.

With those two additions, we have specified the synchronic notion of coherence. Assume that (1) the agent always posts their credences as prices, and (2) the bookie faces the same bounds as the agent. A credence function over $\mathcal{G}$ is coherent iff there is no strategy for such a bookie such that the agent can know beforehand that it'll result in a loss.

For the diachronic version of the argument, everything is exactly the same except we imagine that the whole story gets repeated both before and after a learning event. The agent initially posts odds and the bookie buys and sells tickets with the agent. Then the learning event takes place, where both the bookie and the agent become aware of some new information. We won't put any constraints on what this looks like—we'll treat it as "black box learning", so there's some function from worlds (in the algebra) to the agent's posterior credence function C^+ (Skyrms, 1990; Samet, 1999; Huttegger, 2014).

After the learning event, the agent posts new prices, $C^+(E)$, for all relevant events. The bookie shows up again, and buys and sells more tickets. Adding all the bookie's bets together, if the bookie has a strategy that strictly guarantees a win regardless of what the agent learns (i.e. the agent and bookie are in a position to know that this strategy will net the bookie money), then the agent is diachronically incoherent.

Critical to the diachronic version of the argument is that the agent and the bookie must anticipate the odds and bets that the agent will post, in the relevant possibilities they leave open. Of course, they don't know what they will learn, so they don't know what the agent's posterior credences C^+ are—but they still know conditionals of the form "If the agent is in world w, then the agent will have credence function C^+_w." In effect, this is another way of capturing that the strategy must result in a *foreseeable* loss to the agent—if they update in ways they can't be expected to anticipate (an evil neurosurgeon surreptitiously bonks their brain around), exploiting this doesn't count as a "sure" loss.

Both the non-introspective and unstable agents can be Dutch booked on the sportsbook setup. Consider the following betting strategy. Today, the bookie buys a ticket that pays \$1 in H for \$0.50. Then tomorrow the bookie will adopt the following conditional strategy:

- If the agent posts a price of \$1 on a ticket for H, do nothing
- If the agent posts a price of \$0.50 on a ticket to H, sell two tickets that pay \$1 in H to the agent at \$0.50 apiece

Table 1 shows how these bets play out in each state. The bookie wins money no matter what—note that if the agent posts a price of \$0.50, that means the coin landed tails, so the bookie doesn't need to pay out the bets. The agent has

been booked, and is therefore regarded as incoherent by the sportsbook version of the book argument.

State	H	T
Received from the bookie today	\$0.50	\$0.50
Paid to the bookie today	\$0	\$0
Received from the bookie tomorrow	\$0	\$0
Paid to the bookie tomorrow	\$0	-\$1
Prizes paid to the bookie	-\$1	\$0
Prizes paid to the agent	\$0	\$0
Total	-\$0.50	-\$0.50

Table 1: Money from and to the agent for the sportsbook version of the bets. Positive amounts are money paid by the agent to the bookie, negative amounts are paid by the bookie to the agent. A book is represented by a positive total in each column.

The bookie is choosing a different strategy in the two states, and this may seem like a violation of the symmetric bounds condition. To forestall this objection, note that the bookie is not conditioning his strategy directly on the the state, but rather only on the posted odds of the agent. Of course, the posted odds of the agent pick out, deterministically, the state tomorrow. This is allowed because both the agent and the bookie know this fact ahead of time. The bookie is not using any information the agent did not have. In fact, he is using information that was supplied to him by the agent herself.

This represents an inescapable problem for the unstable agent. However, more might be said about the non-introspective agent. A natural response at this point is to suggest that we've unfairly given the bookie a type of 'second-mover advantage'. Since the bookie can react to the prices the agent posts, shouldn't the agent be able to as well? Perhaps she can update on the credences she is just about to post. She could notice that she is just about to post \$0.50 on H for the second day in a row, and stop herself. She could update on that fact right before posting the odds. Indeed—at least given what we've said she knows about her updating dispositions—this is what she *should* do if she learns her posteriors via looking at her posted odds: like the bookie, she should realize that she'd only offer a price of \$0.50 for bets on H if in fact the coin landed tails. Thus she should condition on this fact, transforming her state of introspection failure (left) to one that satisfies introspection and knows how the coin landed (right):

Thus her 'fair prices' are unstable—they don't survive her announcement of them.[9] We have an sort of 'epistemic tickle' defense against the book (Eells, 1982). Does this mean that our agent is not really subject to a book?

[9] In effect, upon learning what prices she posts our agent becomes informed about what her credences were, and thus (from there-on out) will act in the way Isaacs and Levinstein's 'self-

Proponents of the Reflection princple will have two objections. First, they might claim that this is an escape route very much like the sleight of hand mentioned above, wherein the agent "renormalizes" her credences before announcing her fair prices. This is debatable. The re-normalization case was one where the agent didn't learn anything when they went to announce their fair prices. By stipulation, our case is different: our agent *doesn't* know her fair prices—at least not immediately after the learning experience. If we understand our agent as genuinely learning what her fair prices were (hence what her credences were), then of course she should be allowed to revise her opinions in light of this, just as the bookie is allowed to condition his action on what she announces.

The second objection is more forceful. If the only way our agent can avoid being Dutch booked is by changing her future opinions in this way, then her prior opinions *do* obey Reflection toward her *final* posterior opinions. The Reflection-violations at the intermediate stage seem to be problematically ephemeral. We can put the point as a dilemma: either (i) the agent always revises her non-introspective credences as soon as she starts to act on them, or (ii) she does not. If (i), then Reflection violations will be transitory, and will be removed as soon as she acts—casting doubt on their philosophical importance. If (ii), then our agent *can* indeed be booked by a bookie who's smart enough to realize that she's not properly updating on her knowledge of her own dispositions. Either way, the core idea that Reflection is a rational constraint is more-or-less vindicated.

At this point opponents of the Reflection principle might have a larger objection: the sports book version of the argument is not really suited for considering agents who violate introspection. After all, the agent will end up obeying introspection by fiat, since they must post prices, they cannot post "I don't know," and they can see what prices they're about to post. This might be unfairly biasing the story toward introspective agents. This prompts a move to a different, "offered-bets" version of the Dutch book argument.

3 The offered-bets argument

The offered-bets version of the book argument seems more common in the philosophical literature, and so it might seem a more natural test bed for our question. Like in the sportsbook version, there is an agent and a bookie. There is an algebra of events and tickets just as before. However, instead of the agent posting odds, the bookie must buy or sell tickets with the agent at a price proposed by the bookie and which the agent can either accept or reject. The agent must buy or sell any ticket which she regards as fair or better, calculated using her credences. That means that if the agent's expected monetary return from the sale or purchase is at least $0, then she accepts. Otherwise she rejects. This "fair price" constraint corresponds to the requirement that"the agent posts her credences" in the sportsbook version of the argument.

confident' decision theory prescribes for agents who don't know their own credences (Isaacs and Levinstein, 2024).

Notice that—at least so long as it's mutual knowledge that the agent has these dispositions—then the bookie and agent can infer facts about the agent's credences from which bets she accepts or rejects. Partly because of this, we'll see shortly, that the "symmetric bounds" criteria is much more difficult to interpret in this context, especially in the context of introspection failures. But some aspects of that requirement are easy to understand: the bookie cannot know anything that the agent doesn't. As before, we will say that if the bookie has a strategy that (weakly) guarantees him a win, the agent is incoherent.

In the offered-bets version, the agent does not have to be aware of her credences prior to being offered a bet. Her credences can be a black box to her. She is offered a bet, she consults her black box, and it tells her whether or not to accept. This also means that the bookie can no longer use a strategy that conditions his purchases and sales on the posted odds, since no odds are posted.

However, suppose we allow the bookie to use a *fixed strategy*: he specifies what series of bets he'll offer the agent. Although he can't give different bets (or bets at different prices) in different worlds outright, he *can* take different responses to the agent's actions—if they accept vs. refuse a bet, he can offer different bets in response. If we allow fixed strategies like this, then the bookie can book both the unstable and the non-introspective agents

Here's how. Today, the bookie purchases ten tickets that each pay \$1 in H from the agent at a total price of \$5.00. Tomorrow, the bookie offers to sell just one of those tickets back to the agent at the price of \$0.99 (a "test bet"). The bookie then conditions his next move on the outcome of that offered sale:

- If the sale is accepted, do nothing else. (In this case the coin landed heads, so the other 9 tickets will pay out to the bookie.)
- If the sale is refused, offer to sell all ten tickets the bookie possesses plus ten more that pay \$1 if H at a total price of \$10. This is a fair price by the agent's credences and she accepts. (In this case, the coin landed tails—so the bookie won't have to pay out on the bets.)

As before, this strategy guarantees a win for the bookie—see Table 2.

State	H	T
Received from the bookie today	\$5	\$5
Paid to the bookie today	\$0	\$0
Received from the bookie	\$0	\$0
Paid to the bookie tomorrow	-\$0.99	-\$10
Prizes paid to the bookie	-\$9	\$0
Prizes received from the bookie	\$0	\$0
Total	-\$4.99	-\$5.00

Table 2: Money from and to the agent for the offered bet version. Positive amounts are money paid by the agent to the bookie, negative amounts are paid by the bookie to the agent. A book is represented by a positive total in each column.

The bookie can learn about the agent's credences from seeing what low-stakes bets the agent will accept. Because, in this decision problem, the agent's credences determine what state we are in, this information allows the bookie to condition his betting behavior on the state without relying on that information directly.

But wait—shouldn't we allow the agent to do this as well? Allowing the unstable agent this possibility doesn't matter; she won't adapt. But the non-introspective agent will. In state H the agent would not learn anything, but in state T she would. She could reason this way: "I would only refuse to buy that ticket for \$0.99 in state T, so after conditioning on my own refusal, I now know that I'm in state T."

This would block the bookie's strategy. Now in state T the agent would initially refuse the sale, and that refusal would lead her to revise her credences, leading her to refuse the second sale as well. After refusing the first sale, she *becomes* introspective by observing how she acts.

Indeed, this is what she *should* do. For the bookie's strategy to be a guaranteed win, he needs to *know, from the outset* that she will reject the test bet if and only if she is at T—that's why, after she rejects it, the bookie's strategy of selling bets on T is guaranteed to win him money. But if the bookie knows this, the agent must to. So once she learns (as the bookie does) that she refused the bet, she's in a position to know that she's at T, becoming introspective.

We think this is the correct verdict—it is a genuine way for the agent to avoid this book. But it also suffers from the same problem we observed above: the Reflection failures are ephemeral. Now, they're not *completely* inert: this agent *is* disposed to act differently than a Reflecting agent. After all, at the initial time-step, she's disposed to accept a favorable bet on H—e.g. if offered a bet that paid \$2 if H and $-$\$1 if T, she would accept it at both worlds, while the Reflecting agent would reject it at the T world. More generally, if the agent takes an option that she would take no matter which world she's in, then she doesn't learn anything from her actions and continues to violate Reflection.

But as soon as she *does* take an action that reveals her credences, she will become introspective. That's what happens in the strategy we considered: when the agent rejects the test bet on H, she (along with the bookie) learns that she assigned $C^+(H) = 0.5$, and so infers that she was at the T world. So the trouble is that although non-introspective agent does indeed have different behavioral dispositions than the Reflection-satisfying one, these differences disappear quickly. In the limiting case, if the agent commits to scoring *herself* on a proper scoring rule—one that always incentivizes a probabilistic agent to announce her true credences—then she can always figure out what her credence was just by eliciting her credence from herself, before taking any further actions.

What does this show? We think it shows the importance of the assumption that the agent's credences are known to be reliably linked to her actions—i.e. that she knows she'll take options that maximize expected value relative to C^+. Under this assumption, introspection and (hence) Reflection failures are ephemeral, since the agent in effect has a perfectly-reliable *test* of what her credences are.

The two versions of the book we've looked at so far exploit this test on behalf of the bookie: since he can respond to either the agent's posted prices or actions—which are known to reliably reveal her credences—he can condition his action on what her credences are. As a result, it's not too surprising that an agent in such scenarios can avoid book only if her violations of introspection (and Reflection) are ephemeral.

Of course, you might think that the difference between *no* behavioral difference and *some* behavioral difference is crucial. Under the assumption that the agent knows she'll act reliably on her credences, she can figure out what they are (removing her Reflection failures) easily. But how philosophically-significant this easy-removal is could be debated. We might think of it on analogy with (1) an external test, like a brain-scanner; or instead on analogy with (2) an internal operation, like performing *modus ponens*. On the second analogy, the failure to perform the (mental) action seems like the sort of thing an agent could be rationally criticized for; on the first, it's less clear.

Still, what *is* clear is that the sportsbook and offered-bets version of the Dutch book argument do not allow us to easily assess a situation in which neither the bookie nor the agent can learn from their actions. That raises a question: is there an alternative formulation of the book scenario that is strong enough to book our unstable agent, but which doesn't build in opportunities for the bookie or agent to learn from the agent's actions?

4 Fixed-option books

There is. This formulation is suggested by Dorst et al. 2021; they call them *fixed option* books. In one respect, the fixed option book is more complicated than the first two: the bookie can offer arbitrarily complicated "tickets" to the agent. These tickets can pay any amount of money to either the agent or bookie in any specified state. If the agent regards the ticket as having an expectation equal to or higher than 0 relative to her current credence function, then she must accept the bet. Otherwise, she refuses.

While the bookie has access to complicated tickets in this version, the bookie is restricted in his strategy for offering them. In a fixed option book, the bookie must specify *today* what bets he will offer the agent today and what bets he will offer tomorrow. He can choose different options for each day, but he cannot use the outcome of an offer to alter his strategy. This is the sense in which the options are "fixed". This removes the strategy used by the bookie the previous two versions of the argument, where the bookie conditioned his behavior on either (a) the posted odds or (b) on the agent's response to a test bet.

The unstable agent can be booked with a fixed-option book. The trick is to use "called off" bets where the bet pays zero in particular states. In particular, we will use bets that are called off given things the agent might learn.

The bets are illustrated in Table 3. B_1 is a bet that's partially about the agent's future credences: if the agent updates as she's certain she will, it's just a bet on T; but if she fails to update as she's sure she will (sticking with 0.5

despite being in a heads world, as she does at H_2), it's a disaster. At the initial time, the agent assigns 0 to H_2, so B_1 has positive expected value to the agent given her credences. As a result, she accepts.

State	H_1	H_2	T
B_1	-1	-100	2
B_2	0	7	-3
Total	-1	-93	-1

Table 3: Money from the unstable agent in the fixed-book version of the argument. A book is represented as a negative payout for the agent in each column.

B_2 is the second bet, offered after the agent learns, regardless of what happens. It's a *conditional* bet on H_2, given that the agent is not at H_1—in other words, B_2 is a bet conditional on the agent's posterior. Since H_1 is equivalent to the claim that $\langle C^+(H) = 1\rangle$, it's the same as a conditional bet that says: "if you end up certain of H, the bet's off; if not, it's a bet in favor of H."

If the agent is in H_1, then B_2 is known to give 0, so she will take it (but nets 0 anyways). Meanwhile, if the agent is at T or H_2, she's uncertain between H_2 and T, so it maximizes expected value to take B_2. Thus the strategy that will result from maximizing expected value at each time will be to take B_1 today and take B_2 tomorrow. As the bottom row of the table shows, this results in a sure loss for the agent: in each possibility, she loses at least \$1. The unstable agent is subject to a fixed-option book.

However, there's no equivalent formulation for the non-introspective agent: she is free from fixed option Dutch books. This follows from the fact that the agent "values" her posterior—see §5 below, and Dorst et al. 2021, footnotes 21 and 22. Dorst et al. therefore suggest that "fixed-options" is the correct formal interpretation of the same-bounds condition. If this is correct, then it shows that—properly formulated—there is no genuine Dutch book argument for Reflection in contexts of failures of introspection.

What should we make of this claim? Most authors agree that if an agent is subject to a fixed option book they are incoherent—thus the unstable agent is incoherent.[10] The critical issue is whether an agent who is free from fixed option books, but subject to sportsbook-books or offered-bet-books, is incoherent.

It is worth noting at the outset that the fixed-option book does not apparently miss any obvious cases of incoherence. All synchronic books are fixed option books, and so it will enforce the traditional criteria for synchronic probabilistic coherence. In addition, many failures of diachronic coherence are also fixed option books, as illustrated by a discussion in Dorst et al—for example, if posteriors are certain to be introspective, then Reflection violations are fixed-option bookable. So, it is not *immediately* clear that fixed option books are too narrow.

Whether they are, and whether the fixed-option book is the proper way to understand the fixed bounds condition, is an issue we won't settle here. We

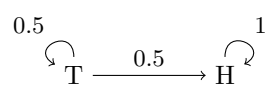
[10] But see Rescorla 2023 for a different, weaker notion of coherence.

suspect that it may be a thorny issue that trades critically on how we think introspection failures might work and what would be an equivalent symmetric condition on the bookie. The critical question: does the fixed option book constrain the bookie too much, giving them access to less than the agent has? How we answer this question may depend on the assumption that's been in the background for much of this discussion: whether the agent is known to reliably act on her credences. We'll return to this in the final section.

5 Rationality and incoherence

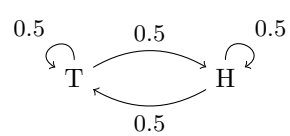
So far, we have demonstrated that the unstable agent is incoherent in all three of the book scenarios, while the non-introspective agent is incoherent on the first two but coherent on the third. Supposing that one of the first two senses is correct—so the non-introspective agent should be judged incoherent—how should she respond to this judgement? She has four options.

Option (1) is to accept the consequences of violating Reflection and be incoherent. She continues on as before, but now perhaps a little less happy with herself. As a reminder, here's how our non-introspective agent's future credences look in each possibility:



We have also already mentioned Option (2): she could make her violation of Reflection transient. She could rewire her dispositions so that she would notice tomorrow that if she's not sure the coin landed heads, then it must have landed tails—hence adjust to $C_T^+(H) = 0$. That yields posteriors:

Option (3) is that the agent could opt to ignore the information about heads and make it so that $C_H^+(H) = 0.5 = C_T^+(H)$. The agent would always stick with 0.5. That yields:



This option also obeys Reflection, but does so at the cost of being less informed.

Option (4) is the least attractive; the agent might arbitrarily alter her credences *today* in order to make them reflect her credences tomorrow. It might at first seem that she could leave her future dispositions the same, but raise her *current* credence in heads to 0.75, to equal her current expectation of her future

credence. But this doesn't work: suppose the agent maintained her knowledge that if heads, she'll become certain of this ($C^+_H(H) = 1$), while if tails she won't move her credence ($C^+_T(H) = C(H)$). Then even if she moves her credence to $C(H) = 0.75$, that will also change her *expectations*—in this case, to $E_C(C^+(H)) = 0.75 \times 1 + 0.25 \times 0.75 = 0.9375$. The only stable place to land would be to *already* be certain that the (fair, not-yet-flipped) coin will land heads! This is Option (4): deciding now that the coin must land heads.

Option (4) is not a serious contender, so there really are only three other alternatives: accept incoherence (Option 1), or save Reflection by choosing (2) or (3). Option (2) is obviously the best among the options, since it involves forming the most accurate credences possible. But what if, for whatever reason, it's not available? (For instance, perhaps she doesn't know she'll act on her credences, so can't infer what they are from her actions—see the next section.)

There's a sense in which Option (3) is worse than (1), despite the fact that the latter violates Reflection. Option (3) involves ignoring some information in favor of obeying Reflection—she's choosing to be less informed in the situation where the coins comes up heads due to the fact that she's inevitably uninformed when it lands tails. This seems needlessly rule-bound—why let the perfect be the enemy of the good? Reflection is supposed to follow from responding optimally to information, rather than being a principle one follows for its own sake.

Indeed, there's a precise sense in which the Option (3) is worse than (1). For any decision problem the agent could face, she thinks that Option-(1) dispositions will lead to at least as good or better outcomes than Option-(3) dispositions would.

Let gamble X be random variable (a function from worlds to numbers), thought of as specifying how much utility you'd receive if you took that gamble in the various worlds you leave open. A decision problem can then be modeled with a set of gambles. When faced with a decision problem $D = \{X_1, .., X_n\}$, the rational thing to do is to take a gamble X_i that maximizes expected value according to your credences C. This quantity is our familiar probability-weighted average of the values of X_i: $\mathbb{E}_C(X_i) = \sum_w C(w) \cdot X_i(w)$.

Say that an agent's current credences C **value** her future credences C^+ iff for any decision problem, the current agent, using her current credences C, would (weakly) prefer to outsource her decisions to her future self using her future credence C^+. (That is, the agent would prefer to costlessly delay a decision.) More precisely: for any $D = \{X_1, ..., X_n\}$, the expected value of using C to take a gamble (choosing a gamble $X^* \in D$ that maximizes $\mathbb{E}_C(X^*)$) is no greater than the expected value of choosing the gamble that maximizes expected value according to C^+.[11] Dorst (2020) and Dorst et al. (2021) propose theories of

[11] Formally, let a **strategy** S be a function from worlds w to options $S_w \in D$, subject to the constraint that if you have the same credences at two worlds then you must choose the same option: if $C^+_x = C^+_y$, then $S_x = S_y$. A strategy S is **recommended** by C^+ iff, at each world w, $S_w \in D$ maximizes expected utility (amongst the X_i) according to C^+_w. The expected value of S is the expected value of following the strategy: $\mathbb{E}_C(S) = \sum_w C(w) \cdot S_w(w)$. Say that C *values* C^+ iff for any decision problem D and any strategy S that C^+ recommends for D, and any other $X \in D$—including the one that C would choose if it decided now—$\mathbb{E}_C(S) \geq E_C(X)$.

rational opinions and updating built around the fact that C can value C^+ even though C does not reflect C^+.

Our non-introspective agent's prior assigns $C(H) = C(T) = 0.5$, and this prior values her posteriors C^+. Why? Because her credences at C^+ are either exactly her credences today (at T) or certain in the correct state (at H); thus she could do no worse by deferring a decision until tomorrow. Compare this to Option (3) of adopting a credence function C^* that assigns 0.5 to heads, come what may. C obeys Reflection toward C^* but not toward C^+. Yet it follows from the fact that C values C^+ (and that C^* is the same as C) that for any decision problem, C would prefer to outsource its decision to C^+ rather than to C^*. Thus whether you prefer to outsource your decision to a credence function (for *every* decision problem) can come apart from whether you obey Reflection toward it.

The failure of introspection is crucial: when credences are introspective, Value is equivalent to Reflection: if C and C^+ are introspective, then C values C^+ iff C reflects C^+ (Huttegger, 2014).

We believe that this discussion shows how there is one sense in which Reflection violations are irrational and another sense in which they are not. If the agent has available to her Option (2)—where she could obey Reflection without loss of information—then she should. However, if she can't do Option (2), and we acknowledge that she is subject to book in some sense, it might still nonetheless be rational for her to violate Reflection. Her bounds force her to choose between two bad choices: incoherence or resistance to information. It is not obvious that incoherence is the worse of the two.

This also shows a broader philosophical point about various properties of ideal Bayesian agents. An ideal Bayesian agent obeys Reflection. They also ideally respond to information (among other things). Placing a constraint on one of these may make the other property no longer ideal.

6 Bounds and Noise

Throughout this discussion we've made the standard assumption that the agent knows she will reliably act on her credences. That assumption has turned out to be surprisingly close to introspection. Compare: if an agent comes equipped with a perfectly-reliable test for what their credences are—like a futuristic brain scanner—it's no surprise that Reflection-failures that result from introspection-failures will be ephemeral. All they need to do is apply the test, figure out their credences, and then they should obey Reflection (cf. Stalnaker, 2019). What we've revealed is that the standard assumption—that people know that their credences are reliably linked to their action—is a way of building in that people have such *do* have such a perfectly-reliable test hanging around with them: all they need to do is act.

This assumption shows up in a different guise when people argue that violations of introspection would lead to 'Moorean' assertions analogous to 'q but I don't believe q'. For example, van Fraassen (2023) claims that if your credence

in q is 0.4 but you don't know what your credences is, you could find yourself saying bizarre things like, 'q is 40%-likely, but I have no idea how likely I think q is.' The oddity here arises because it's presupposed that you know that the first conjunct reliably expresses your probability—in which case you should update on that fact when you say it, falsifying the second conjunct. In contrast, if it's made clear that you *don't* know whether your statement accurately expresses your credence, there is no Moorean oddity: 'I really don't know what I think about q. If I had to guess, I suppose I'd say that I think it's 40%-likely—but I might well think it's 30%- or 50%-likely. I'm not sure.'

This illustrates, we think why it is a commonplace amongst philosophers interested in *failures* of introspection to talk about unreliability and noise in the cognitive system, making it hard to know what you think (e.g. Williamson, 2000; Srinivasan, 2015). What we've established is that this is no accident: in order to make (introspection-based) Reflection failures robust, something must prevent the agent from knowably, reliably acting in ways that clarify what their credences are to themselves.

In closing, we'll look at a couple ways this could go—and what it would mean for the proper account of Dutch books.

6.1 Bounds

One way to get persistent introspection failures is to stipulate that the agent faces sharp bounds on their learning. Perhaps, for some reason, they can't learn from her own behavior (compare Williamson, 2000, §10.6). One could then use the symmetry of bounds condition to argue that the bookie should be prohibited from learning from the agent's behavior as well. This would be another way to justify the fixed-option books.

We think that such a strategy is legitimate but underspecified—it's only a limited defense of Reflection- and introspection-failures. Most sensible agents *can* learn from their own behavior—at the very least, the same way they can learn from the behavior of others. Defending Reflection-failures in this this minmal way is a uncomfortably close to saying, "Reflection failures are permissible for those who can't satisfy Reflection."

Perhaps this inability to introspect might follow from some other, less bespoke bounds that we might think plausible to place on agents. We don't know what this would look like, but think a proposal in the vicinity—*limiting* their ability to learn from their actions without *blocking* it—holds more promise. Let's turn to that.

6.2 Cognitive noise

The idea is based on the hypothesis that our beliefs and actions are pervaded by *cognitive noise*. Plausibly, there is usually (objective) randomness or stochasticity between someone's mental states and the observable ways they act. This is a fairly uncontroversial idea generally—it can be hard to tell whether you're

angry or just tired, since how you'd act (and how you'd respond if you asked yourself) is noisy, hence doesn't perfectly discriminate which state your in.

How could this work for credences? There are a variety of cognitive-science models that implement some version of this idea. The most straightforward says that you have a precise credence t, but when you "elicit" it for action, the actual output is a noisy corruption $t+\epsilon$, where ϵ is a variable with (say) Gaussian noise (Thurstone, 1927; Erev et al., 1994). So if in fact your credence in q is 0.6, and you're offered a bet on q, you take it iff it maximizes expected value according to a probability function that assigns $0.6 + \epsilon$ to q. How likely this is depends on how close to indifferent you are about the bet, and the variance of ϵ—in the limit, as $\epsilon \to 0$, this will correspond to just using your true credence t to make your decision.

A different way of implementing this 'cognitive noise' idea is with the *sampling hypothesis* (Vul et al., 2014; Icard, 2016). The idea—inspired by the computational intractability of exact Bayesian inference, and the plausible algorithms that get around this via approximations—is that if you have credence t in q, you can *draw samples* which are q-possibilities with probability t and $\neg q$-possibilities with probability $1 - t$. If you draw N samples, that's like flipping a t-biased coin N times, leading to an "elicited" credence that follows a Binomial(N, t) distribution. The more samples you draw, the closer this will usually be to t. (There are also more-efficient, though also more-complex, ways of sampling using Markov Chain Monte Carlo.)

Under both the Gaussian and sampling hypotheses, your credences influence your action: your elicited distribution—the thing that directly determines how you act—is *probabilistically* determined by your underlying credence. But at the same time, seeing what you've elicited (or how you've acted) doesn't allow you to be certain of what your credence is, since you're unsure whether this process has been distorted by noise.[12]

What this offers is a sensible picture on which introspection might continue to fail even after you act: when you notice yourself taking an even-odds bet on p, you can't be sure that that means you're at least 50%-confident of p, since you might (say) be 40%-confident but have taken the bet because your elicitation was distorted by noise.

The crucial question: how can we formulate notions of Dutch books for agents with cognitive noise? A Dutch book is supposed to be a *sure* loss; but if there's stochasticity in how the agent will respond to a given option, nothing will be a sure loss even for an agent who is incoherent.

For example, consider the agent who assigns credence 0.75 to heads and 0.75 to tails on a single flip of a coin. Without noise, the bookie can guarantee a win by selling this agent a ticket on heads and a ticket on tails. But, if the agent is subject to cognitive noise, they will sometimes refuse one or both of these bets. So while this agent will still lose money on average, they won't be subject to the

[12]There is a third option. Agents might have initially noisy credences, but once they are elicited the noisy output becomes stipulated as that agent's non-noisy credence in that event. This model of "coherent arbitrariness" has been advocated by (Ariely et al., 2003).

classic notion of book. They will not agree to a series of bets that are certain to lose.

Beyond this concern if the agent only has noisy access to her own credences, how shall we implement the same bounds criterion? Should the bookie only have equally noisy access to the agent's credences? Or should he only be able to respond to her credences after they're elicited—updating on how she acted—as she can?

These are thorny problems, but we don't want to abandon the idea of justifying rational constraints using Dutch books in this context. After all, there seems to be an important difference between (i) an agent who's underlying credences are probabilistic (say, with $C(H) = 0.75$ and $C(\neg H) = 0.25$) but whose elicitations are noisy, and (ii) an agent who's underlying credences are are *not* probabilistic (say, $C'(H) = 0.75$ and $C'(\neg H) = 0.75$), and also has noisy elicitations.

The simplest solution is this: focus on their *underlying dispositions*, screening off the effects of noise. The idea is that we want to know whether their dispositions themselves commit the agent to a sure loss, were it not for the noise that corrupts their elicitations. In other words, we should run the Dutch book argument focusing just on C, rather than it's noisy elicitations.

We assume that both the agent and bookie are informed of what the agent's initial real (unperturbed) credence are and what they would be after various learning scenarios. We must also presume that the agent will in fact act on their unperturbed credences, so that a book is possible.

But the question that must now be asked is what—if anything—is the bookie (and the agent) allowed to learn from how the agent acts, or what credences they post? Of course, if we restrict attention to scenarios the noise doesn't corrupt the agent's actions *and let the agent an bookie know that*, we return to the problem we began with: the agent has a perfectly-reliable test for their credences, so introspection- and Reflection-failures will be ephemeral. We want to focus on the coherence of agents when they know that their dispositions *are* noisy—for this is the scenario in which introspection failures plausibly will be persistent.

On the other hand, if the bookie doesn't know *anything* about the agents credences, he'll have no idea what she'll do (even modulo noise), and so won't be able to book even obviously-incoherent credences like ones that assign $C'(H) = 0.75$ and $C'(\neg H) = 0.75$.

One middle way is to allow the bookie (and agent) to know the agent's credences at the initial time and to know that the agent will act on her credences, *but* to not get any information about what her opinions are later. This restriction would lead to permitting all and only fixed-option Dutch books, for the bookie would not be able to extract any information on the second day—so would in effect have to specify the option-set beforehand. Since a ban on fixed-option Dutch books is equivalent to Value, this would be equivalent to saying that under noise, Valuable violations of Reflection (like our non-introspective agent) are coherent, while non-Valuable ones (like our unstable agent) are incoherent.

While this would pick out the fixed-option Dutch book uniquely, one might worry that it represents a cherry-picked set of knowledge constraints. We don't want to defend or defeat this approach in this paper, but we'd like to make a few points.

First, fixed-option Dutch books don't *require* that the bookie know the agents credences at the initial time.[13] There may be some alternative formulation of a restriction on the bookie's knowledge that also pinpoints fixed-option books.

Second, a ban on fixed-option Dutch books is equivalent to Value. But in some ways, Value is the easier constraint to state and motivate in the context of noise—for it doesn't require the (even metaphorical) existence of a bookie whose knowledge we have to worry about. In the case of noisy credences, what Value says is that, for any decision problem, the agent's prior prefers to outsource their decision *to their underlying dispositions* tomorrow: modulo noise, they expect those future dispositions to do better. If we want to focus on the coherence of the underlying dispositions, this is a natural way to go.

But third, it's reasonable to wonder whether there are just different normative notions of 'coherence' here. It's not obvious whether there's an independently motivated specification of the book scenario that corresponds exactly to Value in the context of noise. Even in the context of noise, the agent (and the bookie) have *some* information about the underlying credences. Perhaps, restricting the notion of book to ignoring this information is unfairly restrictive.

Are there further alternatives? We're not sure. One way to defend Reflection would be to show that, for at least some particular cognitive-noise hypotheses, there *is* a way for the bookie to Dutch book the agent using only legitimate knowledge of their noisiness. The trick in doing this properly will be that the agent themselves has to also have this same knowledge of their noisiness, and update appropriately in response to their own actions. For example, in the above sampling model, if the agent draws n samples and then the bookie and agent both condition on what those samples were, this will generically not tell them exactly what the agent's credences were—introspection (and thus Reflection) failures will remain.

It's and open an interesting question whether there's some other way to make a Dutch book in a case like this—and in particular whether the book will carve a difference between the probabilistic-but-noisy agent and the non-probabilistic agent. The way to try to show it would be to stipulate some simple, tractable noise hypothesis (say, with sampling—since this helpfully keeps the sample space discrete), and then try to construct a Dutch book. We leave exploration of this possibility for future work.[14]

[13] For example, if both the bookie and agent are unsure whether the agent's prior in E is 0.4 or 0.5, but they both know that later the agent will be at least 0.6-confident of E, then the agent is easily fixed-option bookable.

[14] One possibility is to use techniques like those in Hellman 2013; Nascimento 2024 and (separately) De Bona and Staffel 2017, 2018; Staffel 2020 to show that—using normalized bets or accuracy metrics—there is an upper bound on the expected loss from using noisy credences to make decisions. It is an open question whether this approach could draw a sharp line

7 Conclusion

This concludes our dive into the foundations of the Dutch book argument for Reflection. We've canvassed three different types of book scenarios (sportsbook, offered-bets, and fixed-option) and two different ways a synchronically coherent agent can violate Reflection (being unstable or non-introspective).

For all three setups, we've concluded that the unstable agent is incoherent and thus (arguably) irrational, at least in the sense of failing to live up to well-motivated epistemic ideals.

Things were more complicated for the non-introspective agent. When we assume that the agent is known to act reliably on their credences, then she avoids book in the sportsbook and offered-bets formulations only if she learns exactly what her credences are from her actions, and thus ceases to violate Reflection as soon as she acts. Under these assumptions, Reflection violations can be coherent but are always ephemeral—they are edge cases, of mainly theoretical interest. On the other hand, even if the agent learns from her actions, the fixed-option formulation of the book scenario allows her to persistently violate Reflection.[15]

What this discussion repeatedly revealed was that if the agent is known to act reliably on her credences, then there is strong normative pressure toward Reflection. In a way, this isn't surprising: the fixed-option book (and Value) show that the only coherent way to violate Reflection is to fail to be introspective; but if the agent is known to act reliably on her credences, then she can become introspective simply by acting.

We think this observation clarifies why those who defend introspection- (and hence Reflection-) failures must build some version of noise or unreliability into their picture of credences. It also suggests that we should model such unreliability more explicitly than has been done so far, so that we can assess the empirical plausibility, theoretical fruitfulness, and normative contours of such noisy-credence models. The normative plausibility of Reflection-failures hinges on it. We sketched how the *sampling model* provides a natural, tractable hypothesis, and began to explore how to assess its normative standing. This was just a start: future debates over Reflection, introspection, and coherence should proceed by explicitly formulating the types of unreliability or cognitive noise they have in mind, and then exploring the plausible normative constraints for such agents.

Overall, we hope to have clarified the underlying disagreement between those who think that Reflection is normatively sacrosanct and those who think it relies on unrealistic assumptions. These two camps are relying on different background

between probabilistic-but-noisy opinions and (slightly) non-probabilistic opinions. Relatedly, it's worth further exploring the connection between this noise approach and 'random utility' models and their variants (Block and Marschak, 1959; Manski, 1977)—but we'll have to leave that for future work.

[15]Though it's worth mentioning that fixed-option Dutch books and/or Value imply that when you learn the true cell of a partition, you must update by conditioning. So if the agent *does* learn from her actions, she should indeed update on them, even in the fixed-option scenario. The difference is that even if she does so, she avoids fixed-option book even if she doesn't become fully introspective.

conceptions of credence and its connection with action. We think that further exploring these competing conceptions offers a fruitful way forward.[16]

References

Ariely, D., Loewenstein, G. F., and Prelec, D. (2003). "coherent arbitrariness": Stable demand curves without stable preferences. *Quarterly Journal of Economics*, (February):73–105. Citation Key: Ariely2003.

Block, H. D. and Marschak, J. (1959). Random orderings and stochastic theories of response.

Briggs, R. (2009). The Anatomy of the Big Bad Bug. *Nous*, 43(3):428–449.

Christensen, D. (1991). Clever Bookies and Coherent Beliefs. *Philosophical Review*, 100(2):229–247.

Das, N. (2022). Externalism and exploitability. *Philosophy and Phenomenological Research*, 104(1):101–128.

De Bona, G. and Staffel, J. (2017). Graded Incoherence for Accuracy-Firsters. *Philosophy of Science*, 84(2):189–213.

De Bona, G. and Staffel, J. (2018). Why be coherent? *Analysis*, 78(3).

Dorst, K. (2020a). Evidence: A Guide for the Uncertain. *Philosophy and Phenomenological Research*, 100(3):586–632.

Dorst, K. (2020b). Higher-Order Evidence. In Lasonen-Aarnio, M. and Littlejohn, C., editors, *The Routledge Handbook for the Philosophy of Evidence*. Routledge.

Dorst, K. (2023). Rational Polarization. *The Philosophical Review*, 132(3):355–458.

Dorst, K., Levinstein, B., Salow, B., Husic, B. E., and Fitelson, B. (2021). Deference Done Better. *Philosophical Perspectives*, 35(1):99–150.

Eells, E. (1982). Rational Decision and Causality.

Elga, A. (2013). The puzzle of the unmarked clock and the new rational reflection principle. *Philosophical Studies*, 164(1):127–139.

Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological review*, 101(3):519.

Gaifman, H. (1988). A Theory of Higher Order Probabilities. In Skyrms, B. and Harper, W. L., editors, *Causation, Chance, and Credence*, volume 1, pages 191–219. Kluwer.

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

Gallow, J. D. (2021). Updating for Externalists. *Noûs*, 55(3):487–516.

Gärdenfors, P. (1975). Qualitative probability as an intensional logic. *Journal of Philosophical Logic*, pages 171–185.

Harsanyi, J. C. (1967). Games with incomplete information played by

[16] Thanks to Simon Huttegger, Brian Skyrms, Aydin Mohseni, Ben Levinstein, and Bernhard Salow for helpful discussion and comments.

"Bayesian" players, I–III Part I. The basic model. *Management science*, 14(3):159–182.

Hellman, Z. (2013). Almost common priors. *International Journal of Game Theory*, 42(2):399–410.

Huttegger, S. M. (2014). Learning experiences and the value of knowledge. *Philosophical Studies*, 171(2):279–288.

Icard, T. (2016). Subjective Probability as Sampling Propensity. *Review of Philosophy and Psychology*, 7(4):863–903.

Isaacs, Y. and Levinstein, B. A. (2024). Decision Theory without Luminosity. *Mind*, 133(530):346–376.

Kadane, J. B., Schervish, M. J., and Seidenfeld, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, 91(435):1228–1235.

Lasonen-Aarnio, M. (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, 88(2):314–345.

Manski, C. F. (1977). The structure of random utility models. *Theory and decision*, 8(3):229.

Nascimento, L. (2024). *Bounded arbitrage and nearly rational behavior*, volume 77. Springer Berlin Heidelberg.

Rescorla, M. (2023). Reflecting on diachronic Dutch books. *Nous*, 57(3):511–538.

Salow, B. (2018). The Externalist's Guide to Fishing for Compliments. *Mind*, 127(507):691–728.

Samet, D. (1999). Bayesianism without learning. *Research in Economics*, 53:227–242.

Samet, D. (2000). Quantified Beliefs and Believed Quantities. *Journal of Economic Theory*, 95(2):169–185.

Skyrms, B. (1980). Higher Order Degrees of Belief. In Mellor, D. H., editor, *Prospects for Pragmatism*, pages 109–137. Cambridge University Press.

Skyrms, B. (1990). The Value of Knowledge. *Minnesota Studies in the Philosophy of Science*, 14:245–266.

Skyrms, B. (2006). Diachronic coherence and radical probabilism. *Philosophy of Science*, 73(5):959–968.

Srinivasan, A. (2015). Are We Luminous? *Philosophy and Phenomenological Research*, 90(2):294–319.

Staffel, J. (2020). *Unsettled Thoughts: A Theory of Degrees of Rationality*. Oxford University Press, USA.

Stalnaker, R. (2019). Rational Reflection, and the Notorious Unmarked Clock. In *Knowledge and Conditionals: Essays on the Structure of Inquiry*, pages 99–112. Oxford University Press.

Thurstone, L. (1927). A law of comparative judgement. *Psychological Review*, 101(2):266–270.

Uchii, S. (1973). Higher Order Probabilities and Coherence. *Philosophy of Science*, 40(3):373–381.

van Fraassen, B. (1984). Belief and the Will. *The Journal of Philosophy*, 81(5):235–256.

van Fraassen, B. C. (2023). Reflection and conditionalization: Comments on Michael Rescorla. *Nous*, 57(3):539–552.

Vul, E., Goodman, N., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.

Williamson, T. (2008). Why Epistemology Cannot be Operationalized. In Smith, Q., editor, *Epistemology: New Essays*, pages 277–300. Oxford University Press.

Williamson, T. (2019). Evidence of Evidence in Epistemic Logic. In Skipper, M. and Steglich-Petersen, A., editors, *Higher-Order Evidence: New Essays*, pages 265–297. Oxford University Press.