Testing Abductions From Uncertain Evidence

Finnur Dellsén and Borut Trpin

Forthcoming in *Philosophical Quarterly*
(please cite published version when available)

Abstract

Inference to the Best Explanation (IBE) is traditionally conceived of as a rule of inference, in which one infers to the hypothesis that provides the best explanation of one's evidence. But what if some of that evidence is *uncertain*? How, if at all, can the traditional conception of IBE be extended to handle this common class of cases? This paper presents a new general model for investigating rules of inference from uncertain evidence, and then applies this approach to evaluate several different ways of extending IBE to handle uncertain evidence. The model employs computer simulations in which inference rules receive rewards (penalties) for making correct (incorrect) inferences in randomly generated scenarios in which the evidence is uncertain to various degrees. The model allows us to move beyond reliance on intuitive or historical examples in evaluating how well different inference rules handle cases in which the evidence is uncertain.

Keywords: Inference to the Best Explanation, Uncertain Evidence, Epistemic Caution, Evidential Robustness, Defeasible Inference, Computer Simulations

1 Introduction

Standard conceptions of *Inference to the Best Explanation* (IBE) hold that one may defeasibly infer a hypothesis if it offers a better explanation of the totality of one's evidence compared to any other available explanatory hypothesis (e.g. Harman 1965; Lipton 2004; Dellsén 2024). Some prominent philosophers conceive of IBE as a fundamental and autonomous mode of non-deductive inference that justifies belief in, or acceptance of, its conclusions (e.g. Lycan 1985; Harman 1989; McCain & Moretti

2022). On other views of IBE, it complements a more sophisticated form of non-deductive reasoning by acting as a useful heuristic for non-ideal agents. In particular, it has been influentially argued that IBE serves as a heuristic for Bayesian reasoning for non-ideal agents who cannot be expected to assign precise credences to all the propositions involved in Bayesian updating or to constantly update them by conditionalization when they obtain new evidence (e.g. Okasha 2000; Lipton 2001; Dellsén 2018). On either of these conceptions – the fundamentalist and the heuristic conceptions – IBE is a form of defeasible inference from the evidence one has obtained at a given time to the best explanation of that evidence.[1]

But what if one is not certain that some particular (purported)[2] evidence is correct? How, in other words, does IBE deal with uncertain evidence? The short answer is that the standard accounts of IBE referred to above fail to make room for evidence being more or less uncertain in a way that would affect what is inferred from that evidence. These accounts simply take some collection of evidence as given, with no distinction made between what part of that evidence is certain or uncertain, and then tell what to infer from that collection of evidence. This is clearly problematic in cases where the inferability of a given claim depends on the piece(s) of evidence that are uncertain.

Suppose, for instance, that a claim C would best explain evidence $E_1, \dots, E_n$, while an incompatible claim C' would best explain $E_1, \dots, E_{n-1}$, and that E_n is uncertain. What, if anything, should be inferred in that case? To make this more concrete, consider an example: imagine someone hears a loud *t*hud outside (E_t) and sees trash bins *k*nocked over (E_k). The best explanation seems to be that a *r*accoon tipped them over (C_r). But they also think they might have heard a car door *s*lam just before the noise (E_s), which, if accurate, would be better explained by a *h*uman knocking the bins over (C_h). However, they're not sure whether they really heard a car door slam, or imagined it. So what should they infer, as the best explanation of their overall evidence, regarding whether the culprit was a raccoon (as per C_r) or a human (as per C_h)?

To answer questions of this sort, we need to adapt or extend the traditional conception of IBE to handle uncertain evidence such as E_s. But how? This paper explores this issue through the use of computer simulations in which agents obtain combinations of certain

[1] There are also prominent views on which the 'inference'-part of 'IBE' is somewhat misleading, because on these views IBE merely consists in a preference for more explanatory hypotheses within an otherwise standard Bayesian framework for non-deductive reasoning (e.g. Weisberg 2009b; Huemer 2009; Douven 2022; Lange 2022b).

[2] This parenthetical qualification is inserted here to bracket the issue of whether the term 'evidence' is factive. If it is, then it's trivially true (because analytic) that one's evidence is correct, so one cannot be rationally uncertain about evidence. However, it would still not be trivially true that a *purported* piece of evidence is correct, so one could still be rationally uncertain whether a *purported* piece of evidence is correct. Thus, in what follows, readers who take 'evidence' to be factive are invited to mentally insert 'purported' before each occurrence of 'evidence'.

and uncertain evidence, make inferences from that evidence via various IBE-style rules, and are then rewarded or penalized to the extent that they get things right or wrong. This allows us to evaluate different IBE-style inference rules from uncertain evidence in terms of the accuracy of the inferred claims in a range of relevantly different situations.

The remainder of the paper proceeds as follows. In Section 2, we introduce three specific ways of adapting IBE to accommodate uncertain evidence: *Basic IBE* (IBE$_{\text{Ba}}$), *Filtered IBE* (IBE$_{\text{Fi}}$), and *Evidentially Robust IBE* (IBE$_{\text{ER}}$). Section 3 presents the modelling approach we use to evaluate these inference rules, including the setup of our computer simulations, how uncertainty is represented, and the criteria by which inference rules are scored. Section 4 discusses the core assumptions built into our model, as well as the key parameters we systematically vary to test robustness across different scenarios. We report the simulation results in Section 5, highlighting how each rule performs across a range of epistemic contexts and risk profiles. We conclude in Section 6 with a broader reflection on our results and outline avenues for future research.

2 Uncertain Evidence, Inference, and IBE

The general issue of how to handle uncertain evidence has been most thoroughly explored within the Bayesian framework for non-deductive reasoning. Jeffrey (1965) famously proposed a generalization of Bayesian Conditionalization (BC), *viz.* what has become known as Jeffrey Conditionalization (JC). The problem that motivated Jeffrey was, in short, that BC does not countenance the possibility of changing one's subjective probability for an evidential proposition E except by assigning maximal probability to it, i.e. probability one. JC, by contrast, allows for one's subjective probability for any member of some partition of evidential propositions $E_1, \ldots, E_n$ to change to any other probability between zero and one (inclusive). JC then says that one's new probability (P_n) for some hypothesis H, $P_n(H)$, should be a weighted sum of one's previous conditional probabilities (P_o) of H given each E_i, where the weight attached to each E_i is one's new probability in E_i: $P_n(H) = \sum_{i \leq n} P_n(E_i)\, P_o(H \mid E_i)$.[3]

In a recent paper, Trpin & Pellert (2019) show how JC may be combined with an IBE-inspired rule for updating subjective probabilities that has been developed and defended by Douven (2013, 2022).[4] Since Douven calls his rule EXPL, we will refer to this combined rule as JC$_{\text{EXPL}}$. Trpin and Pellert argue, based on computer simulations

[3] Note that for the rule to be applicable, the so-called *rigidity* condition needs to hold, i.e. $P_n(H \mid E_k) = P_o(H \mid E_k)$ for all k. See Weisberg (2009a) and Schwan & Stern (2017) for a discussion.

[4] The rule Douven develops and defends was originally suggested, and then criticized, by van Fraassen (1989), in an influential critique of Inference to the Best Explanation.

similar to those reported on below, that JC$_{\text{EXPL}}$ is preferable to JC in a number of situations, roughly because it generally leads to more accurate probability distributions than JC.

While JC$_{\text{EXPL}}$ offers a promising approach to probabilistic updating on uncertain evidence, our main concern in this paper is with more traditional conceptions of IBE on which it involves *inferring* some claim C rather than merely updating one's probability in C. Put differently, we will be focusing on accounts of IBE in which one ends up accepting or believing C, rather than merely assigning some new subjective probability to C. This makes our task harder insofar as we cannot directly avail ourselves of Jeffrey's influential proposal for handling uncertain evidence, e.g. in the way Trpin and Pellert do in JC$_{\text{EXPL}}$. However, addressing these harder cases is important, since several popular conceptions of IBE, grounded in scientific practice, involve agents accepting or believing, rather than just updating their probability for, the best explaining hypothesis.

Consider, then, how these traditional inferential conceptions of IBE may be adapted or extended to handle situations in which some of one's evidence is uncertain. The most flat-footed way of doing so would perhaps be to make no distinction between evidence that is more and less (un)certain, treating all evidence alike. On this version of IBE, one simply infers to the claim C that best explains all of one's (relevant) evidence, regardless of whether, and the extent to which, one is uncertain about some of that evidence. In what follows, we will refer to this as *Basic IBE* (IBE$_{\text{Ba}}$). In our earlier example of a knocked-over trash bin, this rule would warrant an inference to a human being the culprit (C_h) rather than a raccoon (C_r), even though one is uncertain about the car door slam (E_s) relative to which the former counts as a better explanation than the latter.

A slightly more sophisticated way of extending inferential conceptions of IBE would be to filter out the evidence that's too uncertain. So, in a case where some pieces of evidence $E_1, \ldots, E_{k-1}$ count as certain, while some other pieces of evidence $E_k, \ldots, E_n$ count as uncertain, one then simply infers to the claim that best explains $E_1, \ldots, E_{k-1}$ and ignores $E_k, \ldots, E_n$. We'll call this *Filtered IBE* (IBE$_{\text{Fi}}$).[5] IBE$_{\text{Fi}}$ thus warrants inferring that the raccoon tipped over the trash bin (C_r) in our earlier example, since the uncertain car door slam (E_s) is simply ignored.

On an intuitive level, it seems clear that neither Basic IBE (IBE$_{\text{Ba}}$) nor Filtered IBE (IBE$_{\text{Fi}}$) could be quite right. Consider the uncertain piece of evidence E_s again. IBE$_{\text{Ba}}$

[5] Filtered IBE assumes that any given proposition can be classified as *certain* or *uncertain* for the purposes of a given inference. This bifurcation of evidential propositions into certain and uncertain ones is much cruder than, say, assigning a degree of (un)certainty to each proposition, as JC (and JC$_{\text{EXPL}}$) effectively does via its probability assignments to the evidence $P_n(E_i)$. However, this crudeness is arguably appropriate when constructing a rule of defeasible inference that is meant to be used by agents who aren't in a position to engage in more nuanced forms of reasoning, such as credence updating by JC (or JC$_{\text{EXPL}}$).

treats it as if it were just as credible as other, certain, pieces of evidence. That seems misguided in so far as, from the agent's own perspective, uncertain pieces of evidence may well have been misperceived. IBE_{Fi} errs in the other direction by treating E_s as if it isn't evidence at all. That can't be quite right either as uncertain evidence, while perhaps not fully reliable, still provides some information that should not simply be ignored. This motivates, at an intuitive level for now, the introduction of some third alternative on which uncertain evidence such as E_s is treated differently both from the having of uncertain evidence and from not having that evidence at all. But what might such an alternative look like?

In a recent paper, Dellsén (2025) proposes a rule of this kind, *Evidentially Robust IBE* (IBE_{ER}). IBE_{ER} is designed to handle uncertain evidence in a way that only requires the agent to classify evidence as either certain or uncertain. Very roughly, IBE_{ER} licences an inference to some claim C just in case, regardless of whether any given uncertain piece of evidence is taken into account or not, C is implied by the best explanation of that evidence. To illustrate, consider again the example from the introduction. Here, IBE_{ER} does not licence inferring either that a raccoon tipped over the trash bin (C_r, which explains the certain part of one's evidence), or that a human did so (C_h, which best explains all evidence). Rather, IBE_{ER} instead licences inferences to those claims that are implied by both C_r and C_h – for instance, that some animal caused the disturbance (C_a), and not, say, a gust of wind.

In fleshing this idea out more precisely, Dellsén defines a notion of an *open evidential combination* as any set of evidential propositions which includes all certain such propositions, and some, none, or all uncertain evidential propositions. Put differently, an open evidential combination is any set that has as a subset the set of all *certain* evidential propositions, $\mathbb{E}_c$, and is itself a subset of *all* (certain and uncertain) evidential propositions, $\mathbb{E}$. Formally, $\mathbb{E}^o_k$ is an open evidential combination if and only if $\mathbb{E}_c \subseteq \mathbb{E}^o_k \subseteq \mathbb{E}$. Intuitively speaking, these combinations of evidence are 'open' in the sense that one hasn't settled on a single one of them as the definitive evidential input into one's inference to the best explanation.

With this notion in place, IBE_{ER} can now basically be thought of as a version of IBE in which one hedges one's bets when some of one's evidence is uncertain by inferring only what the best explanations of *each* open evidential combination have in common. More precisely, IBE_{ER} states that, if C is implied by each hypothesis H_i that provides a better explanation than any other available hypothesis of some open evidential combination $\mathbb{E}^o_k$, then one may infer C. IBE_{ER} thus licences inferences to claims that are robust across different open evidential combinations. Since such claims will generally be less informative than those that can be inferred via IBE_{Ba} or IBE_{Fi}, IBE_{ER} in effect trades informativeness for epistemic caution, permitting only those inferences that are robust across all evidential combinations.[6]

Dellsén motivates this particular way of handling uncertain evidence within an IBE-style framework by considering how it handles a hypothetical case due to van Fraassen (1980, 19–20), and the historical case of Einstein's reaction to numerous experiments that appeared to demonstrate effects of 'ether drift', which would have falsified his special theory of relativity. The former suggests that IBE$_{\mathrm{ER}}$ is intuitively plausible (at least in some cases), and the latter that it corresponds to actual scientific practice (at least in some cases). With that said, a more systematic investigation into the merits and demerits of IBE$_{\mathrm{ER}}$ seems to be called for. Preferably, such an investigation should be less reliant on intuitions about hypothetical cases and the specifics of historical case studies (which may or may not be representative of the general class of cases in which IBE-style reasoning is called for).

The computer simulations we report on in what follows are designed to fill this lacuna. We constructed a model that allows us to simulate how well agents using IBE$_{\mathrm{Ba}}$, IBE$_{\mathrm{Fi}}$, and IBE$_{\mathrm{ER}}$ succeed in making inferences that are both true and informative under a range of different conditions and specifications.

3 The Modelling Setup

The model in which we propose to test the various inference rules outlined above may be seen as an extension of the models that have been used in the extant literature to simulate IBE-inspired versions of BC and JC (Douven 1999, 2013, 2022; Trpin & Pellert 2019). In each round of simulation in these models, an agent receives evidence in the form of results of coin flips, and then updates their subjective probabilities in various possible biases of the flipped coin. In a similar manner, we consider agents who, in each round of simulation, receive evidence of (apparent) results of coin flips and then *infer* to a particular bias of that coin, or to the disjunction of several such biases, using one of the rules discussed above. Our overall concern is to figure out which rule performs best, in the sense of producing the most accurate beliefs about the actual biases of the coins.

While our model involves probabilistic elements, for instance in how coin biases and

[6] It is worth noting that IBE$_{\mathrm{ER}}$ should probably not be understood as a procedure that agents must strictly and consciously follow in all cases of abductive reasoning. Understood in this way, IBE$_{\mathrm{ER}}$ would simply be far too cognitively demanding in many cases, especially when there are many pieces of uncertain evidence. After all, given n pieces of uncertain evidence, there are 2^n evidential combinations that the agent would in some sense have to take into account. Instead, IBE$_{\mathrm{ER}}$ may be more charitably understood as a recommended method for reasoning abductively when and to the extent that the agent is able to keep track of these evidential combinations, such as when the number of uncertain pieces of evidence happens to be low.

evidential reliability are represented, it is important to note that the agents themselves are not modelled as probabilistic reasoners. Probabilities appear only in the description of the environment and in the coin biases posited by each hypotheses under considerations, providing a precise and transparent framework for comparing qualitative IBE-style rules. The rules themselves operate on categorical classifications of evidence as certain or uncertain and on explanatory fit, and involve the agents making categorical inferences about which hypotheses are correct.[7]

A key question is how to model the possibility of uncertain evidence in these situations. Our approach is as follows. In each round of simulation, the coin is flipped a fixed number of times (e.g. ten), producing a sequence of heads and tails. These are the *facts* about the coin landings, not the agent's (possibly uncertain) evidence. The agent then receives a sequence of outcomes each tagged with a degree of *evidential confidence* – that is, a probability indicating how confident the agent is that a particular flip landed heads or tails. For example, the agent might be highly confident that the first flip was heads but highly uncertain about the second, perhaps due to impaired observation (e.g. poor lighting).

To capture the possibility that uncertain evidence may be misleading, we assume that what the agent registers as evidence in such cases may not fully reflect the actual facts about the coin landings. In particular, we adopt an assumption of calibrated evidential uptake, according to which evidential confidence corresponds to the chance of correct registration of evidence. On this assumption, if an agent's evidential uncertainty regarding some evidential proposition E_i is some probability x, then the chance that E_i is true is also x. To illustrate with a simple example, suppose that a coin is biased to always land heads, such that it will in fact produce the sequence H, H, H, $\ldots$ (using H as a shorthand for heads). When the agent receives evidence about the first coin flip, with some evidential confidence attached to it, for example 90%, then there will be a 90% chance that the agent's evidence is factive, and a 10% chance that it's not. Thus there is a 10% chance that the evidential proposition to which the agent attaches a 90% evidential confidence is nevertheless false.

This means that the agents in our model may sometimes 'misregister' the actual result of coin flips. However, the probability of correct registration of the evidence increases with higher evidential confidence. Importantly, whether or not the perceived outcome matches the fact, the agent's evidential confidence in that perception remains unchanged. As a result, misleading evidence can occur despite high evidential confidence, although such cases are relatively rare under our assumption of calibrated evidential uptake. Of

[7] We briefly discuss the possibility of fully probabilistic reasoners who follow expected accuracy maximisation in Section 6.

course, agents should ideally rely directly on the facts, but we assume that these are not directly accessible. Instead, agents must reason on the basis of what they take to be the outcome, together with how certain they are of it, where this certainty reflects the likelihood that their evidence is veridical, though it offers no guarantee. We regard this as a plausible idealization for modelling agents whose evidential access is limited but systematically connected to the facts.

Although our model thus employs degrees of evidential confidence, the IBE rules that we aim to investigate operate with a cruder classification of evidence into *certain* or *uncertain*. To bridge this gap, we assume that each agent operates with thresholds of evidential confidence which determine which evidential propositions get classified as certain and uncertain for the purposes of a given inference. This allows the model to incorporate probabilistically nuanced inputs while preserving the simplicity of reasoning processes that rely on qualitative distinctions between types of evidence. For instance, suppose the agent's thresholds are such as to categorize any evidential proposition as uncertain if their evidential confidence is between 0.5 and 0.8. If the agent is 0.7 confident that the coin landed heads in some toss, they treat that as uncertain evidence. If their evidential confidence had instead been 0.9, they would have treated it as certain. And if an agent were only 0.3 confident of, say, heads, they would treat it as uncertain evidence of tails (we assume here that tails is the negation of heads, so it would be assigned $1 - 0.3 = 0.7$ confidence). See Table 1 for an illustrative example.

Toss #	**True Outcome**	**Evidential Confidence**	**Registered Evidence**	**Classification (threshold = 0.8)**
1	H	0.90	H	Certain
2	H	0.75	H	Uncertain
3	H	0.60	T	Uncertain
4	H	0.95	H	Certain
5	H	0.80	H	Certain
6	H	0.55	T	Uncertain
7	H	0.70	H	Uncertain
8	H	0.55	T	Uncertain
9	H	0.85	H	Certain
10	H	0.90	H	Certain

Table 1: Illustration of uncertain evidence uptake. In this case, the true outcome is always Heads (H), but the agent's subjective evidence may differ depending on their confidence per toss. Subjective evidence is determined stochastically. For instance, with 0.6 confidence in Heads, there's a 60% chance they correctly register it as Heads and a 40% chance they misregister it as Tails. Depending on their confidence, the agent classifies each observation as *Certain* (above 0.8) or *Uncertain* (between 0.5 and 0.8).

The simulated agent then receives a batch of evidence in the described way and categorizes each item as certain or uncertain, depending on their evidential confidence thresholds for each. They then apply one of the inference rules to infer a claim C about the coin's bias. Suppose, for instance, that the agent knows that there are only three candidate coins: one that is two-tailed, one that is two-headed, and one that is fair. In the example from Table 1, the agent infers as follows depending on which rule they employ. If using $\mathrm{IBE}_{\mathrm{Ba}}$, the agent infers that the coin is fair, since that best explains the totality of their evidence (which includes both H and T outcomes). If using $\mathrm{IBE}_{\mathrm{Fi}}$, the agent infers that the coin is two-headed, since that best explains the certain part of their evidence (which includes only H outcomes). Finally, if using $\mathrm{IBE}_{\mathrm{ER}}$, the agent infers that the coin is either two-headed or fair, since that's implied by every best explanation of an open evidential combination (note that no open evidential combination includes only T outcomes).

In each case, there is a factual answer as to whether the inferred claim is indeed correct. Our model includes a scoring system in which an agent using some particular inference rule is rewarded for inferring correct claims, and penalized for inferring incorrect claims. For instance, if the coin is in fact two-headed (as in in the Table 1 scenario), $\mathrm{IBE}_{\mathrm{Ba}}$ is penalized (for incorrectly inferring that the coin is fair), while both $\mathrm{IBE}_{\mathrm{Fi}}$ and $\mathrm{IBE}_{\mathrm{ER}}$ are rewarded. However, as we discuss below, our scoring system is designed to be such that $\mathrm{IBE}_{\mathrm{ER}}$ receives a smaller reward in this type of case due to the fact that its inference is less informative: the disjunction 'two-headed *or* fair' is logically weaker than 'fair', so it is more likely to be correct by default. We return to the issue of how to construct an appropriate scoring system below.

4 Assumptions and Evaluation Criteria

In addition to these basic features, our model makes a number of more specific assumptions that call for some discussion. Some of these assumptions are *fixed*, in the sense that they won't be altered as we run different variations of our simulations. Other assumptions will be adjusted as we go through these variations; we refer to them as *parametric* assumptions (and to the collection of mutually exclusive such assumptions as a parameter). As we shall see, some of the more significant results we obtain are stable across such parametric assumptions, which suggests that they are not mere artifacts of these assumptions.

Let us start by discussing the most notable fixed assumptions of the model. First, we will assume that the agent in question is confronted with a set of mutually exclusive and exhaustive hypotheses, such as the three options in the aforementioned example (two-tailed, two-headed, or a fair coin). Thus, it is not part of the agent's epistemic conundrum to figure out what sort of hypotheses could potentially explain the evidence, or whether there are unconceived alternatives that provide even better explanations than those the agent has already considered. We assume that these issues are already settled before the agent makes an inference from uncertain evidence, since it is not our aim here to test different approaches to this well-known problem (see e.g. van Fraassen 1989; Lipton 1993; Dellsén 2021).

Second, we will also assume throughout that the agent is always able to correctly identify the hypothesis that would provide the best explanation of some evidential propositions. In particular, we follow van Fraassen (1989) in assuming that the agent identifies the best explanation by comparing the observed ratio of heads to tails with the biases associated with each candidate hypothesis, and selecting the hypothesis (or hypotheses in case of a tie) whose predicted bias most closely matches this ratio. This approach, more generally known as the maximum likelihood method, treats explanatory goodness as a function of fit between the evidence and the explanatory hypothesis. For example, if the agent's evidence contains seven heads and three tails, then among candidate biases for heads of 0.3, 0.5, 0.7, and 0.9, the hypothesis assigning a 0.7 probability to heads would count as the best explanation.

We adopt this method not because it captures all intuitive dimensions of explanatory goodness, but because it offers a well-understood, principled, and computationally tractable way of implementing IBE in simulations (see Glass 2007, 2012 for a discussion of its merits and limitations). It is particularly helpful that we have a good way of determining explanatory goodness at least in this limited context because it is highly contested how to measure explanatory goodness in general.[8] This means that there are important aspects of abductive reasoning as it is actually used in science – having to do with how a hypothesis' explanatory goodness is determined by factors such as its parsimony, explanatory scope, unifying power, and so forth – that our model intentionally abstracts away from. For this reason, the results we discuss below should not be taken to provide any insights on these aspects of abductive reasoning.

Other fixed assumptions of the model relate specifically to the fact that our setup involves inferences from observations of coin tosses to hypotheses about how the tossed coin is or isn't biased. While this is a somewhat artificial type of case, of course,[9]

[8] For some influential probabilistic attempts, see Good (1968); Schupbach & Sprenger (2011); Glass (2012), but note that all of these attempts have been criticized. See also Lange (2022a) for an argument against the possibility of any such general measure.

[9] Note, for example, that it's assumed in these simulations that the coin can only land heads or tails, and that they relevant agent knows this. As a result, the agent's attitudes with regard to evidential propositions can only go wrong in one way, i.e. by the agent thinking that a coin landed heads (tails) when in fact it landed tails (heads). This simplification makes the case somewhat unlike many other cases, e.g. observations of the colour of some object, in which an observation can go wrong in a number of different

these cases nevertheless have several advantages over other cases for testing IBE-style inference rules. First of all, such cases fit nicely into a computational model due to being analytically clean, conceptually transparent, and computationally tractable. Second, as we'll see, these cases are also very flexible with respect to the sorts of parametric assumptions that we'd like to vary in our simulations, such as the amount of evidence obtained and the number of hypotheses in play. Finally, the fact that coin tossing cases have been used previously in similar contexts, e.g. by Douven (2013, 2022) and Trpin & Pellert (2019), is itself a reason to use similar cases here, because it makes it easier to compare our results to this previous body of work.

Let us now turn to the parameters of the model. One important parameter is the amount of evidence that the agent obtains before making their inference. In our cases, this translates to the number of times the coin is flipped before inferring to some bias of that coin. Since coin tosses are an extremely low-cost way to gather evidence, in reality one would presumably flip the coin hundreds or thousands of times for any case in which the stakes are even somewhat high. Doing so could swamp any potential error due to randomly misleading evidence and lead agents to make correct inferences in almost every case. However, in the default setup of our simulations we will assume a low number of coin tosses, *viz.* ten, so as to have our cases resemble what we take to be a common feature of the contexts in which IBE is frequently used, such as scientific and legal contexts. In these contexts, gathering additional evidence is often either not possible, or so costly that the agent is prepared to make an inference from relatively meagre evidence. We will, however, also consider cases in which the agent obtains significantly more evidence – specifically, twenty-five, fifty, or one hundred outcomes of coin tosses.

Another parameter is the number of alternative hypotheses that are 'in play', i.e. the number of candidate biases of the flipped coin from which the agent selects one as the best explanation of some evidence. In principle, this number could be infinite, but we follow both Douven (2013, 2022) and Trpin & Pellert (2019) in using finite, and rather modest, sets of hypotheses in our simulations. Specifically, in different variations on our simulations we consider agents choosing between three, seven, and eleven hypotheses, corresponding to low, medium and relatively high granularity in the hypothesis space. In each case, the potential biases will be at equal intervals, and in each case they'll range from zero (0%) to one (100%) bias for heads. For example, when there are seven hypotheses in play, the candidate biases for heads are 0, 1/6, 1/3, 1/2, 2/3, 5/6, and 1.

A third parameter concerns the evidential confidence thresholds by which agents categorize evidence as certain or uncertain. As we've intimated, the inference rules we

ways (e.g. a green object may be observed to be blue, or red, or yellow, etc).

consider do as such not dictate any specific evidential confidence thresholds. For this reason, we explore different threshold settings in our simulations. We will, however, assume that the lower threshold for uncertainty is always 0.5, so that there is no 'uncertainty gap' between one's uncertainty in E_i and one's uncertainty in $\neg E_i$. Put differently, the probability of an evidential proposition E_i always maps exactly onto one and only one of four categories: (i) certainty in E_i, (ii) uncertainty in E_i (iii) uncertainty in $\neg E_i$, or (iv) certainty in $\neg E_i$. These might correspond, for instance, to the following probability ranges for E_i: (i) $[0.8, 1]$; (ii) $[0.5, 0.8)$; (iii) $(0.2, 0.5]$; and (iv) $[0, 0.2]$. Whether this is the best partition is one of the issues under investigation here. In particular – and noting that, due to the symmetry, we need only specify the cases for (i) – we consider threshold schemes where the agent treats E_i as certain if their evidential confidence in E_i lies in one of the following ranges: $[0.5, 1]$, $[0.6, 1]$, $[0.7, 1]$, $[0.8, 1]$, or $[0.9, 1]$, corresponding to increasingly higher thresholds for classifying an observation as certain evidence.

A fourth and final parameter concerns what rewards and penalties are dispensed in each round of simulation, i.e. the structure of the scoring system. The most straightforward (but as we shall see, overly simplistic) scoring system would be one on which each rule gets +1 unit of rewards for inferring a true claim, and -1 unit for inferring a false claim. However, one might also want to see how the rules perform relative to reward systems that aren't perfectly balanced in this way, i.e. systems that are appropriate for more or less risk averse agents. For instance, one might want to penalize each rule for getting things wrong more than one rewards them for getting things right. In what follows, we therefore also consider how our inference rules perform relative to reward systems in which the reward for being right, r, is not necessarily equal to the penalty for being wrong, p.

Another point about the aforementioned simple scoring system is that it does not take into account the extent to which a true inferred claim is *informative*. For instance, if H_1 is a true explanatory hypothesis, then a rule would receive a reward of +1 for inferring H_1, but it would also receive +1 for inferring $H_1 \vee H_2$, and similarly for $H_1 \vee H_2 \vee H_3$, etc. Clearly, our rules should not receive the same rewards for inferring such highly disjunctive claims as for their non-disjunctive, or just less-disjunctive, alternatives. Intuitively speaking, this is because longer disjunctions carry less information about the bias of the coin and are therefore automatically more likely to be correct. Note that this matters for our purposes because IBE_{ER} is designed to often infer to a less informative claim C, such as a disjunction of two or more biases, as a way of hedging its bets in cases of evidential uncertainty. The alternatives against which we test it below, by contrast, do not sacrifice informativeness for epistemic security in this way. So the aforementioned straightforward/simplistic reward system would seem to stack the deck in favor of IBE_{ER}.

To get around this issue, we devised a scoring system which dispenses rewards in proportion to the informativeness of its true conclusions. Specifically, it rewards inferential rules according to how many of the false hypotheses in play are ruled out by the inferred claim. For example, suppose (as in Table 1) that three different coin biases are live options – 0, 0.5, and 1 bias for the coin to land heads – and that 1.0 is the correct bias (i.e. the coin is two-headed). Now, in such a case, a rule like IBE$_{\text{ER}}$ infers the disjunction of the last two biases (0.5 or 1), thereby ruling out one of two false hypotheses. In this round, IBE$_{\text{ER}}$ is then rewarded $1/2 \times r = r/2$ units. By contrast, the rule IBE$_{\text{Fi}}$ that infers the (non-disjunctive) true hypothesis that the bias for heads is 1.0 would be rewarded $2/2 \times r = r$ units. As this illustrates, inference rules are rewarded for making more informative inferences, provided that the inferred claims are true. If the inferred claim is false, e.g. as when the rule IBE$_{\text{Ba}}$ infers that the coin is fair, then it is penalized p units. This system thus nicely avoids stacking the deck in favor of inference rules that, like IBE$_{\text{ER}}$, are designed to sometimes take a lesser epistemic risk.

More formally, our approach to scoring inferential rules is as follows: if an inferred claim C contains the true explanatory hypothesis, and the number of false live hypotheses is m and C contradicts k of these hypotheses, then the rule from which C was inferred is rewarded $k/m \times r$ units, where r is the reward. (Since we'll sometimes assume that $r = p = 1$, where p is the penalty for the incorrect inference, this sometimes simplifies to k/m.) In the special case, where C just is the true hypothesis (and not, for example, a disjunction of it with one or more false hypotheses), we have that $k = m$, so the reward is simply r (or, with the aforementioned simplification, 1). If, however, the inferred claim C does not contain the true hypothesis, the rule is penalized p units. This reflects the principle that true inferences vary in epistemic value by informativeness, but all false inferences are equally bad, however cautious, specific, or inclusive they may be.

5 IBE, Uncertainty, Risk, and Complexity

We now turn to reporting on the results of our simulations, in which an agent – using one of the three versions of IBE described above (see Section 2) – attempts to infer the bias of a coin (selected randomly from a predefined set) on the basis of a sequence of coin flips, some of which are treated as uncertain. The agent then receives a reward or penalty depending on whether their inference was true or false and how informative the inferred claim was, as explained above (see Section 4). To recap, the parameters of our model are:

(i) the amount of evidence obtained before inferring – in our case, the number of coin landings (ten, twenty-five, fifty, or one hundred);

(ii) the number of live alternative hypotheses – in our case, the number of equally spaced out biases for heads from zero to one (three, seven, or eleven);

(iii) the probability threshold separating what the agent considers certain from uncertain evidence (0.5, 0.6, 0.7, 0.8 or 0.9); and finally,

(iv) how much the rule is rewarded for making a correct inference versus penalized for making an incorrect inference (thirteen possible combinations of rewards and penalties with values of zero, one, two, and ten).[10]

In total, this means that we simulated 780 ($4 \times 3 \times 5 \times 13$) parameter combinations. For each of these setups, we simulated inference performance across situations in which each of the coin biases included in the live hypothesis set was the actual bias of the coin. That is, when the hypothesis space contained three possible coin biases, we simulated how well the inference rules performed relative to three actual coin biases (i.e. two-tailed, fair, and two-headed), and similarly for the seven- and eleven-bias configurations. This resulted in an average of seven distinct coin biases (because (3+7+11)/3=7) per parameter combination. We therefore ran 5,460 simulations with 1,000 rounds for each setup to ensure robust and reliable results. We then computed the average net scores (rewards minus penalties) associated with each inference over these 1,000 rounds.

The mechanics of the simulations are summarised in Figure 1. Each run begins with parameter selection, followed by generation of coin toss sequences with varying evidential confidence. The selected inference rule is applied to the categorised evidence, the result is scored, and the process is then repeated until it's reached 1,000 rounds.

Given the large number of possible parameter combinations, we focus here on a set of representative and philosophically illuminating cases. The aim is to not overwhelm the reader with an exhaustive catalogue of all simulated cases by instead highlighting the patterns that emerge across a range of salient parameter combinations. These cases are not cherry-picked: interested readers can inspect or extend the full simulation set using the publicly available code.[11]

As a starting point, we ran a baseline set of simulations with moderate values for each parameter: ten coin tosses per round, seven possible coin biases (ranging from

[10] We exclude the option where both the reward and the penalty have the value zero because such scenarios are trivial – all inference rules would receive the same score. We also exclude the options where the reward and penalty are both two or ten because the relative scores would be the same as when the reward and penalty are both one.

[11] The simulation code, implemented in Python, is available at `https://github.com/philosophy-simul/ibe_er_simulations`.

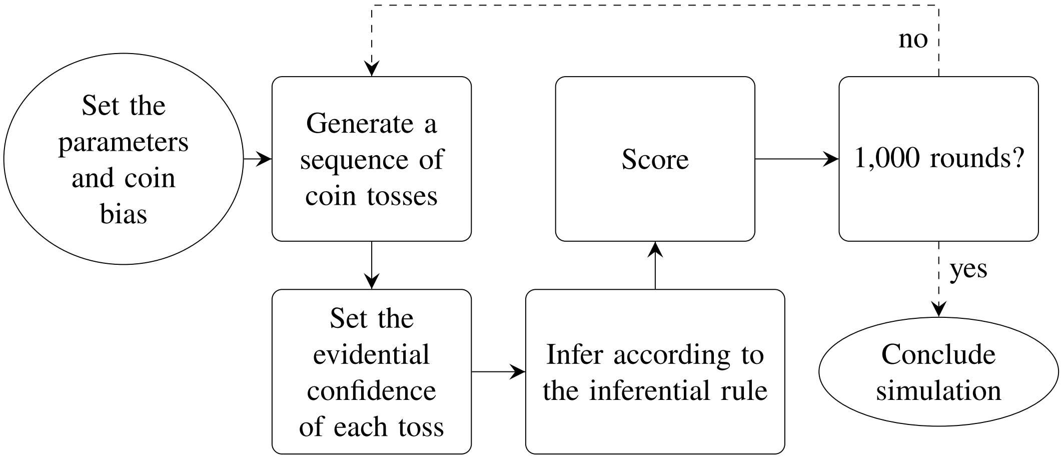
no
Set the parameters and coin bias
Generate a sequence of coin tosses
Score
1,000 rounds?
yes
Set the evidential confidence of each toss
Infer according to the inferential rule
Conclude simulation

Figure 1: Flowchart of the computer simulations (dashed lines are conditional).
Alt text: A flowchart demonstrating how the simulation proceeds. In each of the 1,000 rounds, a sequence of coin tosses and their evidential confidence is generated, from which an inference according to an inferential rule follows. The inference is scored and the cycle repeats unless 1,000 rounds have passed when it concludes.

0 to 1.0 bias for heads), a balanced scoring system in which fully informative correct inferences were rewarded (+1) and incorrect ones penalized (–1) (see Figure 2). Under these relatively neutral conditions, IBE$_{\text{ER}}$ consistently outperformed other rules in terms of overall score. Its advantage was particularly clear when the threshold for treating evidence as certain was neither too high nor too low, i.e. when the agent classified evidence with a probability between 0.5 and 0.8 as uncertain. These results support the idea that evidential caution, in the sense of avoiding overreliance on uncertain evidence without disregarding it completely, can lead to more accurate inference in contexts where uncertain evidence is common and epistemic stakes are balanced.

Let us now turn to the influence of each parameter so as to systematically assess how different inferential strategies fare under changing epistemic conditions.

First, let's consider the amount of evidence available to the agent (see Figure 3). As expected, increasing the number of coin tosses (from ten to one hundred) leads to improved performance across all inference rules. However, the relationship between evidence volume and reliability varies markedly by rule. For IBE$_{\text{Ba}}$, which considers all available evidence regardless of the evidential confidence attached to it, performance on average improves with more tosses. However, the variability also increases significantly because when this rule goes wrong, it tends to go very wrong with more evidence. This reflects IBE$_{\text{Ba}}$'s vulnerability to misleading evidence, which increases rapidly with an greater amounts of uncertain evidence.

In contrast, IBE$_{\text{ER}}$, which considers uncertain evidence more cautiously than IBE$_{\text{Ba}}$,

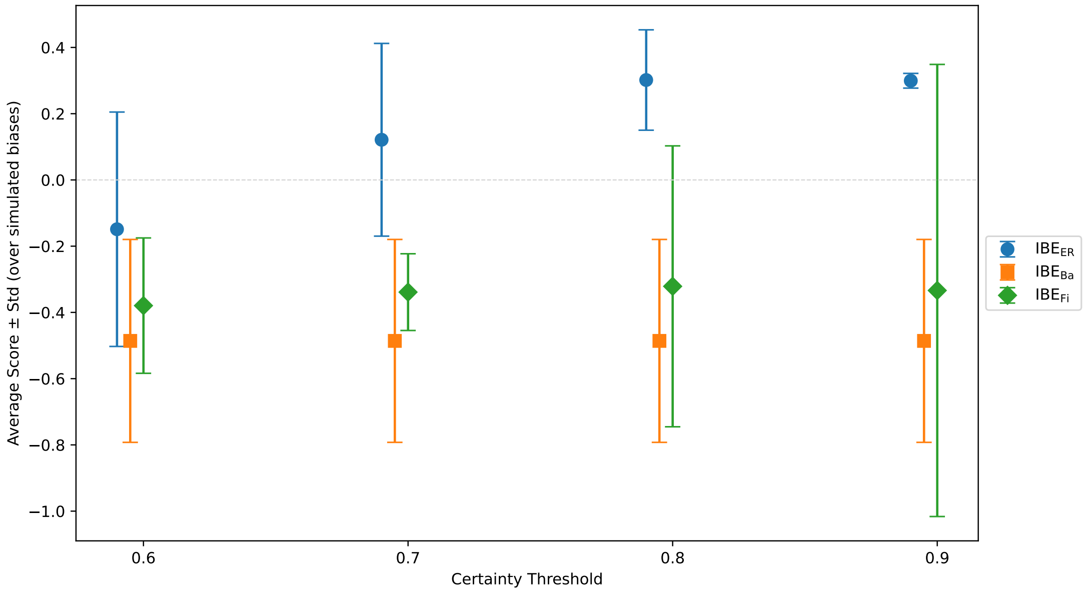

Figure 2: Performance in terms of average score (± standard deviation) by certainty threshold under baseline conditions (ten coin tosses, balanced reward and penalty, seven live hypotheses). Note that $\mathrm{IBE}_{\mathrm{Ba}}$ is threshold-insensitive because it disregards whether evidence is (un)certain.
Alt text: A graph showing average scores and standard deviations of inferential rules relative to different certainty thresholds. The evidentially robust version of IBE tends to perform the best.

exhibits markedly different pattern. With more evidence (certain or uncertain) and a stricter certainty threshold, we see both an improvement in its average performance, and less variability in its scores. This suggests that being cautious about which evidence to consider also helps stabilize inference via $\mathrm{IBE}_{\mathrm{ER}}$ when more evidence becomes available.

$\mathrm{IBE}_{\mathrm{Fi}}$, which completely disregards uncertain evidence, exhibits a pattern that falls in between those exhibited by $\mathrm{IBE}_{\mathrm{Ba}}$ and $\mathrm{IBE}_{\mathrm{ER}}$. $\mathrm{IBE}_{\mathrm{Fi}}$'s performance does improve with more evidence, but its variability in performance does not decrease in the same way as for $\mathrm{IBE}_{\mathrm{ER}}$. If the evidential confidence threshold is too low (so that the rule does not disregard much evidence), $\mathrm{IBE}_{\mathrm{Fi}}$ can be misled more easily, and if it is too high, there may not be enough evidence, which leads to greater variability (note, in particular, its performance when the lower threshold for evidential confidence is 0.9 and there are just ten coin tosses).

Next, we turn to the number of live hypotheses the agent from which the agent is choosing the best explanation – specifically, whether the possible coin bias could be one of three, seven, or eleven equally spaced coin biases between zero and one for the coin to land heads (see Figure 4). As expected, increasing the number of live hypotheses makes

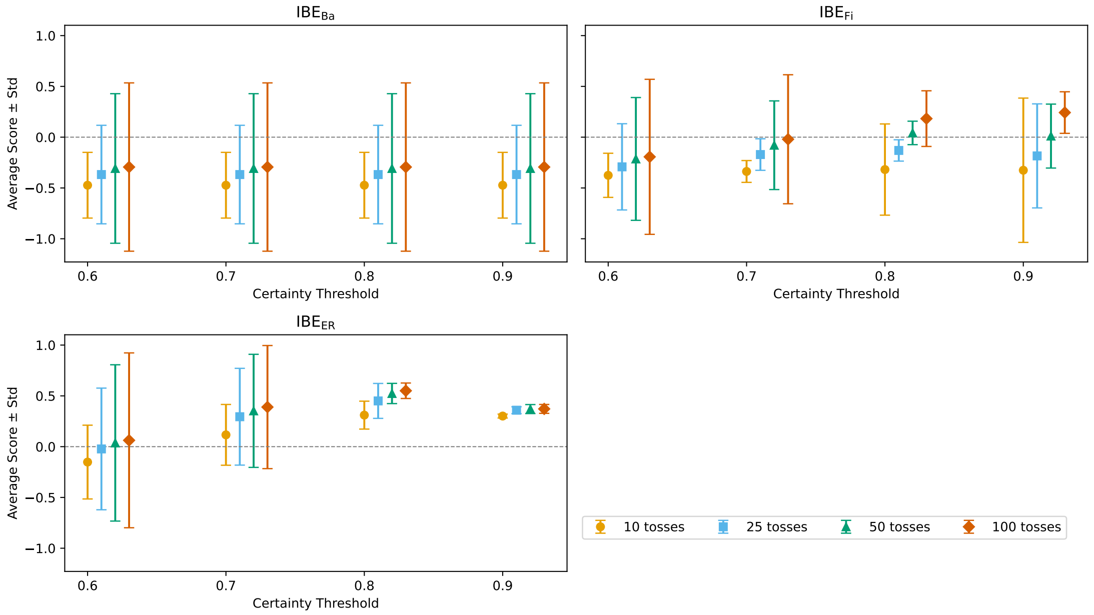

Figure 3: Average scores by amount of evidence (number of tosses) in the default setting (seven hypotheses, balanced penalties and rewards of one unit).
Alt text: A graph showing average scores and standard deviations of inferential rules relative to the number of coin tosses. IBE Basic does worst with lowest averages and widest variability. IBE Filtered does better on average but remains widely varied. IBE ER does the best with highest averages and lowest variability when certainty threshold is high.

the inference task more challenging, so the average scores decrease across the board. Nevertheless, the performances of the inference rules interact with this complexity in revealing ways.

When the hypothesis space is small (for instance, just three options), both IBE$_{Ba}$ and IBE$_{Fi}$ perform well and often outperform IBE$_{ER}$. In such low-complexity settings, the task resembles a low-risk gamble, where drawing specific conclusions from limited evidence could yield relatively high rewards. In this case, being less cautious and including more uncertain evidence tends to pay off.

As the number of hypotheses increases, however, this strategy becomes less effective. For more complex inference tasks (with seven or eleven live hypotheses), IBE$_{ER}$ consistently outperforms its competitors, especially when configured with a higher certainty threshold. By avoiding premature commitment to a single specific bias and taking into account that some of the evidence may be unreliable due to its uncertainty, IBE$_{ER}$ seems to be better equipped to handle ambiguity and reduce the risk of erroneous inferences, even though the inferred claims are less informative (and thus less rewarding by our scoring system). A plausible lesson is that it pays off to infer less informative claims when one's evidence is ambiguous and the inferential task is moderately complex.

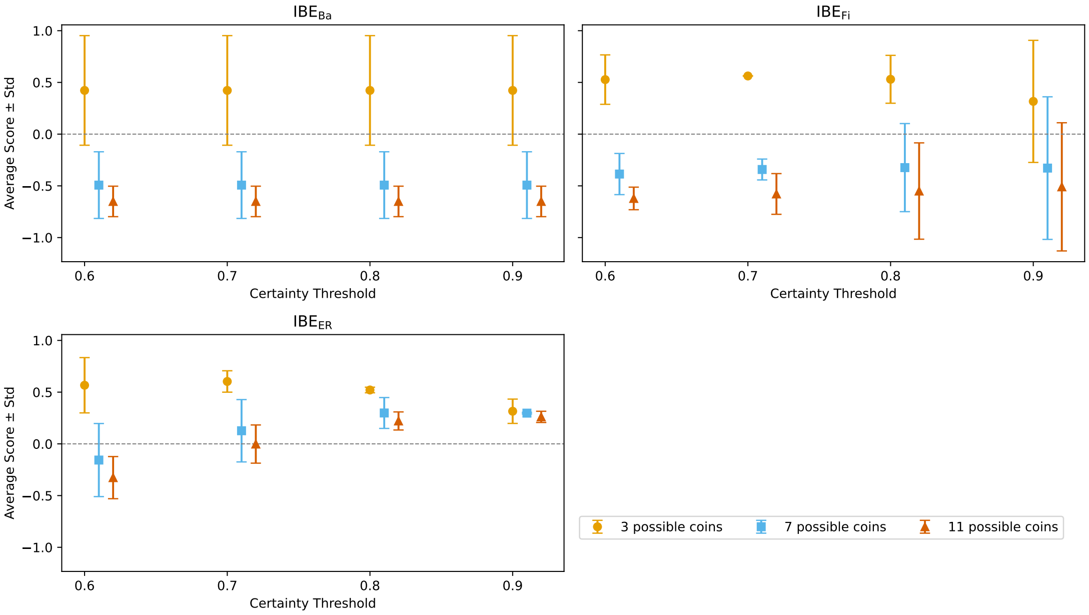

Figure 4: Performance by hypothesis space size (25 coin tosses, balanced penalties and rewards of one unit).
Alt text: Graph showing average scores and standard deviations for different inferential rules by relative to different sizes of hypothesis space sizes (three, seven, or eleven possible coins). There is no obvious best rule.

The certainty threshold, *i.e.* the threshold in terms of evidential confidence above which the agent counts a piece of evidence as certain, also shapes outcomes in revealing ways. This threshold does not have a uniformly optimal setting; rather, how well inferential rules with different certainty thresholds perform in our simulations depends on the complexity of the inferential task and the rewards/penalties for making correct/incorrect inferences. In cases where evidential confidence is low, there are many live hypotheses, or the cost of error is high, higher certainty thresholds help IBE$_{ER}$ maintain its advantage by excluding misleading evidence. But in low-risk or simpler settings, where guessing is relatively cheap and often correct, such cautiousness is less helpful. Here, inference rules that are more aggressive, in the sense of inferring to a single specific hypothesis, tend to perform better. Similarly, if one lowers the certainty threshold, IBE$_{ER}$ also becomes less cautious and performs better. This underscores the adaptive value of tuning evidential thresholds to context rather than treating them as fixed parameters. This can be easily seen when we compare the results of low risk to those of high risk in Figure 5.

Further, to assess how the rules respond to differences in the cost of error, we can adjust the reward-penalty structure in our scoring system. In low-risk, high-reward scenarios (e.g. no penalty, +10 reward, abundant evidence, few live hypotheses), aggressive inference tends to be the better strategy. Here, rules like IBE_{Ba} or, especially, IBE_{Fi} gain an edge by making bolder, more informative inferences that are often correct, where IBE_{ER} tends to often infer to a disjunction of multiple possible biases and therefore scores lower. Conversely, in high-risk scenarios, where incorrect inferences are penalized more heavily than correct ones are rewarded (e.g. -10 penalty to 0 reward), and the task is complex (little evidence, many live hypotheses), IBE_{ER}'s caution pays off. It substantially outperforms other rules by avoiding costly mistakes. See Figure 5 for two illustrative cases.

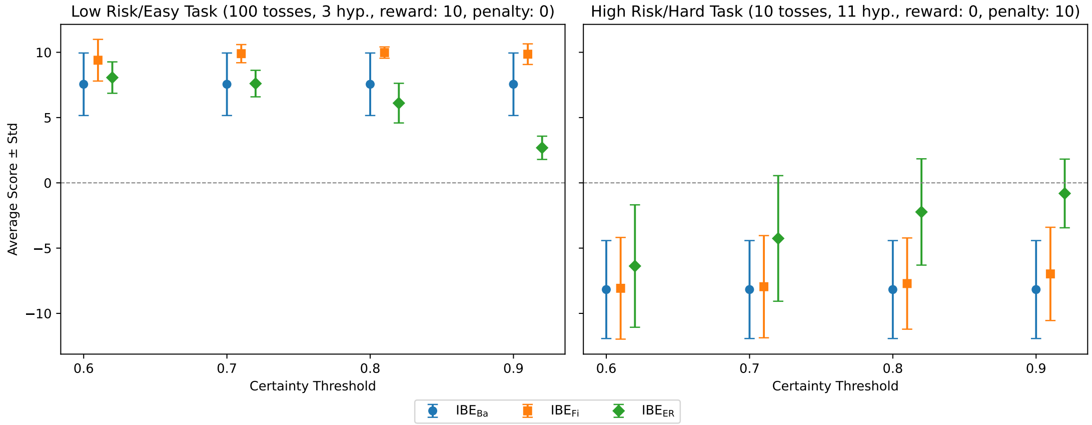

Figure 5: Scores across risk/reward conditions and varying complexities of the inferential task.
Alt text: A graph of two subplots comparing two setups: low risk and easy task vs. high risk and hard task. IBE Basic and IBE Filtered are on par or better to IBE ER in the easy variant, but outperformed by IBE ER in the harder variant.

Taken together, these simulation results reveal how IBE_{Ba}, IBE_{Fi}, and IBE_{ER} have agents infer differently across a wide range of epistemic conditions. No single of these inference rules dominates across all parameter specifications, but clear patterns emerge: IBE_{ER} fares best in cases characterized by moderate to high uncertainty about the evidence, complexity in the inferential task, and higher risk of error. IBE_{Ba} and IBE_{Fi} sometimes perform better than IBE_{ER}, but primarily when there is the inferential task is simple, there is considerably less uncertainty about the evidence, or when the risk of error is low compared to the reward for getting things right. These findings underscore the importance of tailoring the employment of inferential rules to the case at hand. With this in mind, we now turn from quantitative performance to a broader reflection on what these results suggest about IBE and evidential uncertainty.

6 Concluding Remarks

This paper has developed and applied a new modelling approach for investigating inference rules under conditions of evidential uncertainty. We simulated three different ways in which Inference to the Best Explanation (IBE) can be extended to handle uncertain evidence, and compared the performances of these rules under a variety of different conditions. Our simulations show that, in a wide range of conditions, Evidentially Robust IBE (IBE_{ER}) outperforms its two most straightforward alternatives, Basic IBE (IBE_{Ba}) and Filtered IBE (IBE_{Fi}). This provides argumentative support for IBE_{ER} beyond the sort of intuitive cases and scientific case studies that have thus far been used to motivate it (Dellsén 2025).

With that said, our simulations also reveal how the performances of different IBE-style inference rules depend not on their internal logic alone, but also on the context in which they are deployed. In simple, low-risk settings with little evidential uncertainty, IBE_{Ba} and IBE_{Fi} often perform as well as, or even better than, IBE_{ER}. This strongly suggests that there may be no universally correct IBE-style inference rule from uncertain evidence; rather, different situations call for different ways of inferring from uncertain evidence. This context-dependence of epistemic success accords with the programme of *ecological rationality*, which evaluates heuristics and rules by how well they are adapted to the informational structure and payoff profile of their environments rather than by context-insensitive standards alone (*e.g.* Gigerenzer & Goldstein 1996; Todd & Gigerenzer 2012). It also connects with recent work in epistemology that develops explicitly environment-relative accounts of rational inquiry and rule evaluation (*e.g.* Douven 2020, 2025; Thorstad 2024). Although our aim has not been to develop environment-adaptive rules, our results nonetheless align with and lend support to ecological approaches by demonstrating that the performance of inference rules depends systematically on features of the environments in which they are applied.

More broadly, our results speak to an important philosophical point about how inference rules should be evaluated. It makes little sense to evaluate an inference rule solely by its performance in ideal situations; it must also be assessed by how it performs under the kinds of evidential imperfections that actual agents face. Inference rules like IBE_{ER}, which build uncertainty directly into their structure, demonstrate that it is possible to reason well without treating uncertainty as noise to be ignored, eliminated, or idealized away. Rather, uncertainty itself can be the object of principled epistemic treatment. To do this, we must ask which forms of inference are most reliable in the evidential conditions in which agents typically operate.

We'll end by making two concrete suggestions for future research on this topic.

First, our model operates with evidence that is 'binary' in the sense that each piece of evidence is one of two possible outcomes, *viz.* a coin landing heads or tails. One might worry that this simplicity of our model limits the generality of our results, in so far as much evidence is significantly richer in content, *e.g.* by being one of several, perhaps even infinitely many, possible outcomes. While many real-world inference tasks are based on 'binary' evidence of this kind,[12] it is also undeniable that this choice imposes a limitation. A more general approach might explore richer forms of evidential uncertainty. For example, instead of binary coin setups, one could simulate inferences over biased dice with multiple faces. It would be interesting to see whether $\mathrm{IBE}_{\mathrm{ER}}$ would continue to dominate in contexts where the inferential task is complex, the evidence is noisy, and mistakes are costly; or whether more assertive strategies such as $\mathrm{IBE}_{\mathrm{Ba}}$ and $\mathrm{IBE}_{\mathrm{Fi}}$ would do better than in the simulations on which we've reported here.

Second, while the target of our investigation was IBE-style inferences, the broader modelling approach we have employed here is not restricted to inference that proceed via the identification of the best explanations of evidence. Our modelling approach can therefore be extended to other kinds of inferences from uncertain evidence, provided they generate categorical outputs or can be paired with a suitable decision procedure that provides a categorical output. For instance, one could use our model to investigate the performance of probabilistic belief updating schemes which were developed with uncertain evidence in mind, such as Jeffrey conditionalization (Jeffrey 1965) or its explanationist generalisation (see Trpin & Pellert 2019), given an appropriate decision-theoretic procedure for translating beliefs into categorical inferences. An obvious way of doing so would be to consider all possible inferences one can draw (e.g. that a single coin is used in a simulation or a disjunction of several of them) and infer to the option with the maximum expected accuracy. The resulting inference rule could then be compared, in much the same way we have done here, to other inferences rules from uncertain evidence, such as $\mathrm{IBE}_{\mathrm{ER}}$.

Notes

Both authors contributed equally to the research and writing of this paper.
Finnur Dellsén would like to thank the Iceandic Research Fund (Grant No. 228526-051) for research funding while working on this project.

[12] Medical tests often return only a positive or negative result. Witness testimony may consist in a simple affirmation or denial. Scientific instruments sometimes just register a detection event or fail to do so. In such cases, agents are faced with the task of selecting the best explanation from multiple hypotheses based on discrete, fallible observations. Modelling inference from binary evidence is therefore not just a theoretical convenience, but a way to capture an important class of real-world epistemic problems.

Borut Trpin would like to thank the Slovenian Research and Innovation Agency (ARIS, research project No. J6-60107 and research core funding No. P6-0144) for their kind support of his research, and to LMU Munich for supporting his research stay at the University of Iceland in May 2025.

Thanks to the audiences at “Formal Models of Scientific Inquiry” (Ruhr University Bochum, 2025) and EPSA 2025, for their helpful comments on earlier versions of this paper.

Finnur Dellsén
University of Iceland, Iceland
University of Inland Norway, Norway
University of Oslo, Norway

Borut Trpin
University of Ljubljana, Slovenia
University of Maribor, Slovenia

References

Dellsén, F. (2018) ‘The Heuristic Conception of Inference to the Best Explanation’, *Philosophical Studies*, 175: 1745–66.

Dellsén, F. (2021) ‘Explanatory Consolidation: From ‘Best’ to ‘Good Enough’’, *Philosophy and Phenomenological Research*, 103/1: 157–77.

Dellsén, F. (2024) *Abductive Reasoning in Science*, Cambridge: Cambridge University Press.

Dellsén, F. (2025) ‘Inferring to the Best Explanation from Uncertain Evidence’, *Philosophy of Science*. Forthcoming.

Douven, I. (1999) ‘Inference to the Best Explanation Made Coherent’, *Philosophy of Science (Proceedings Supplement)*, 66/S3: S424–35.

Douven, I. (2013) ‘Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation’, *Philosophical Quarterly*, 63/252: 428–444.

Douven, I. (2020) ‘The ecological rationality of explanatory reasoning’, *Studies in History and Philosophy of Science Part A*, 79: 1–14.

Douven, I. (2022) *The Art of Abduction*, Cambridge, MA, MIT Press.

Douven, I. (2025) ‘Adaptive Updating: Ecological Rationality Meets Reinforcement Learning’, *Minds and Machines*, 35/32: 1–23.

Gigerenzer, G. & Goldstein, D. G. (1996) ‘Reasoning the fast and frugal way: models of bounded rationality.’, *Psychological Review*, 103/4: 650–69.

Glass, D. H. (2007) ‘Coherence Measures and Inference to the Best Explanation’, *Synthese*, 157: 275–96.

Glass, D. H. (2012) ‘Inference to the Best Explanation: Does It Track Truth?’, *Synthese*, 185: 411–27.

Good, I. J. (1968) ‘Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor’, *The British Journal for the Philosophy of Science*, 19/2: 123–43.

Harman, G. (1965) ‘The Inference to the Best Explanation’, *The Philosophical Review*, 74/1: 88–95.

Harman, G. (1989) *Change In View: Principles of Reasoning*, Cambridge, MA: MIT Press.

Huemer, M. (2009) ‘Explanationist Aid for the Theory of Inductive Logic’, *British Journal for the Philosophy of Science*, 60/2: 345–75.

Jeffrey, R. (1965) *The Logic of Decision*, New York: McGraw-Hill.

Lange, M. (2022a) ‘Against Probabilistic Measures of Explanatory Quality’, *Philosophy of Science*, 89/2: 252–67.

Lange, M. (2022b) ‘Putting Explanation Back Into “Inference to the Best Explanation?”’, *Noûs*, 56/1: 84–109.

Lipton, P. (1993) ‘Is the Best Good Enough?’, *Proceedings of the Aristotelian Society*, 93: 89–104.

Lipton, P. (2001) ‘Is Explanation a Guide to Inference? A Reply to Wesley Salmon’, in G. Hon & S. Rakover (eds.) *Explanation: Theoretical Approaches and Applications* 93–120. Dordrecht: Kluwer Academic Publishers.

Lipton, P. (2004) *Inference to the Best Explanation*, 2nd edn. London: Routledge.

Lycan, W. G. (1985) ‘Epistemic Value’, *Synthese*, 64: 137–164.

McCain, K. & Moretti, L. (2022) *Appearance and Explanation*, Oxford: OUP.

Okasha, S. (2000) ‘Van Fraassen’s Critique of Inference to the Best Explanation’, *Studies in the History and Philosophy of Science*, 31: 691–710.

Schupbach, J. N. & Sprenger, J. (2011) ‘The Logic of Explanatory Power’, *Philosophy of Science*, 78/1: 105–127.

Schwan, B. & Stern, R. (2017) ‘A Causal Understanding of When and When Not To Jeffrey Conditionalize’, *Philosophers’ Imprint*, 17.

Thorstad, D. (2024) *Inquiry Under Bounds*, Oxford: OUP.

Todd, P. M. & Gigerenzer, G. (2012) *Ecological Rationality: Intelligence in the World*, Oxford: OUP.

Trpin, B. & Pellert, M. (2019) ‘Inference to the Best Explanation in Uncertain Evidential Situations’, *British Journal for the Philosophy of Science*, 70/4: 977–1001.

van Fraassen, B. C. (1980) *The Scientific Image*, Oxford: Clarendon.

van Fraassen, B. C. (1989) *Laws and Symmetry*, Oxford: Clarendon.

Weisberg, J. (2009a) ‘Commutativity or Holism? A Dilemma for Conditionalizers’, *The British Journal for the Philosophy of Science*, 60: 679–860.

Weisberg, J. (2009b) ‘Locating IBE in the Bayesian Framework’, *Synthese*, 167: 125–143.