Higher-order Probabilism

2024-05-15

Abstract. Rational agents are often uncertain about the truth of many propositions. To represent this uncertainty, it is natural to rely on probability theory. Two options are typically on the table, precise and imprecise probabilism, but both fall short in some respect. Precise probabilism is not expressive enough, while imprecise probabilism suffers from belief inertia and the impossibility of proper scoring rules. We put forward a novel version of probabilism, higher-order probabilism, and we argue that it outperforms existing alternatives.
Keywords: Probabilism; Imprecise probabilities; Evidence; Probability; Belief inertia; Bayesian networks; Proper scores.

1 Introduction

As rational agents, we are uncertain about the truth of many propositions since the evidence we possess about them is often fallible. To represent this uncertainty, it is natural to rely on probability theory. Two options are typically on the table: precise and imprecise probabilism. Precise probabilism models an agent's state of uncertainty (or credal state) with a single probability measure: each proposition is assigned one probability value between 0 and 1 (a sharp credence). The problem is that a single probability measure is not expressive enough to distinguish between intuitively different states of uncertainty rational agents may find themselves in (§2). To avoid this problem, a *set* of probability measures, rather than a single one, can be used to represent the uncertainty of a rational agent. This approach is known as imprecise probabilism. It outperforms precise probabilism in some respects, but also runs into problems of its own, such as belief inertia and the impossibility of defining proper scoring rules (§3).

To make progress, this paper argues that the uncertainty of a rational agent is to be represented neither by a single probability measure nor a set of measures. Rather, it is to be represented by a higher-order probability measure, more specifically, a probability distribution over parameter values interpreted as probabilities. The theory we propose is not mathematically novel, but addresses many of the problems that plague both precise and imprecise probabilism (§4 and §5). It also fares better than existing versions of probabilism when the probability of multiple propositions, dependent or independent, is to be assessed (§6 and 7). Many of the examples in this paper are about coin tosses, but in the final two sections, we will

also discuss a couple of legal examples as a hint to the broader applicability of higher-order probabilism.

2 Precise probabilism

Precise probabilism holds that a rational agent's uncertainty about a proposition is to be represented as a single, precise probability measure. Bayesian updating regulates how the prior probability measure should change in light of new evidence that the agent learns. The updating can be iterated multiple times for multiple pieces of evidence considered successively. This is an elegant and simple theory with many powerful applications. Unfortunately, representing our uncertainty about a proposition in terms of a single, precise probability measure runs into a number of difficulties.

Precise probabilism fails to capture an important dimension of how our fallible beliefs reflect the evidence we have (or have not) obtained. A couple of stylized examples featuring coin tosses should make the point clear. Here is the first:

> **No evidence v. fair coin** You are about to toss a coin but have no evidence about its bias. You are completely ignorant. Compare this to the situation in which you know, based on overwhelming evidence, that the coin is fair.

On precise probabilism, both scenarios are represented by assigning a probability of .5 to the outcome *heads*. If you are completely ignorant, the principle of insufficient evidence suggests that you assign .5 to both outcomes. Similarly, if you know for sure the coin is fair, assigning .5 seems the best way to quantify the uncertainty about the outcome. The agent's evidence in the two scenarios is quite different, but precise probabilities fail to capture this difference.

And now consider a second scenario:

> **Learning from ignorance** You toss a coin with unknown bias. You toss it 10 times and observe *heads* 5 times. Suppose you toss it further and observe 50 *heads* in 100 tosses.

Since the coin initially had an unknown bias, you should presumably assign a probability of .5 to both outcomes if you stick with precise probabilism. After the 10 tosses, you again assess the probability to be .5. You must have learned something, but whatever that is, it is not modeled by precise probabilities. When you toss the coin 100 times and observe 50 heads, you learn something new as well. But your precise probability assessment will again be .5.

These examples suggest that precise probabilism is not appropriately responsive to evidence. Representing an agent's uncertainty by a precise probability measure can fail to track what an agent has learned from new evidence. Precise probabilism assigns the same probability in situations in which one's evidence is quite different: when no evidence is available about a coin's bias; when there is little evidence that the coin is fair (say, after only 10 tosses); and when there is strong evidence that the coin is fair (say, after 100 tosses). In fact, analogous problems also arise for evidence that the coin is not fair. Suppose the rational agent starts with a weak belief that the coin is .6 biased towards heads. They can strengthen that belief by

tossing the coin repeatedly and observing, say, 60 heads in 100 tosses. But this improvement in their evidence is not mirrored in the .6 probability they are supposed to assign to *heads*.[1]

These problems generalize beyond cases of coin tossing. Suppose that, given a certain stock of evidence, the probability of A is greater than that of B. Further, suppose that the acquisition of new evidence does not change the probabilities. Admittedly, something has changed in the agent's state of uncertainty: the quantity of evidence on which the agent can assess whether A is more probable than B has become larger. And yet, this change is not reflected in the precise probabilities assigned to A and B.[2]

3 Imprecise probabilism

What if we give up the assumption that probability assignments should be precise? Imprecise probabilism holds that a rational agent's credal stance towards a hypothesis is to be represented by a set of probability measures, typically called a representor $\mathbb{P}$, rather than a single measure P. The representor should include all and only those probability measures which are compatible with the evidence (more on this point later).[3] It is easy to see that modeling an agent's credal state by sets of probability measures avoids some of the shortcomings of precise probabilism. For instance, if an agent knows that the coin is fair, their credal state would be represented by the singleton set $\{\mathrm{P}\}$, where P is a probability measure that assigns .5 to *heads*. If, on the other hand, the agent knows nothing about the coin's bias, their credal state would be represented by the set of all probabilistic measures, since none of them is excluded by the available evidence. Note that the set of probability measures does not represent admissible options that the agent could legitimately pick from. Rather, the agent's credal state is essentially imprecise and should be represented by means of the entire set of probability measures.

So far so good. But, just as precise probabilism fails to be appropriately evidence-responsive in certain scenarios, imprecise probabilism runs into similar difficulties in other scenarios.

> **Even v. uneven bias:** You have two coins and you know, for sure, that the probability of getting heads is .4, if you toss one coin, and .6, if you toss the other coin. But you do not know which is which. You pick one of the two at random and toss it. Contrast this with an uneven case. You have four coins and you know that

[1] Another problem for precise probabilism is known as *sweetening* (Hare, 2010). Imagine a rational agent who does not know the bias of the coin. For precise probabilism, this state of uncertainty is represented by a .5 probability assignment to heads. Next, the agent learns that the bias towards heads, whatever the bias is, has been slightly increased, say by .001. Intuitively, the new information should leave the agent equally undecided about betting on heads or tails. After sweetening, the agent still does not know much about the actual bias of the coin. But, according to precise probabilism, sweetening should make the agent bet on heads: if the probability of heads was initially .5, it is now be slightly above .5.

[2] The distinction here is sometimes formulated in terms of the *balance* of the evidence (that is, whether the evidence available tips in favor a hypothesis or another) as opposed to its *weight* (that is, the overall quantity of evidence regardless of its balance); see Keynes (1921) and Joyce (2005).

[3] For the development of imprecise probabilism, see Keynes (1921); Levi (1974); Gärdenfors & Sahlin (1982); Kaplan (1968); Joyce (2005); Fraassen (2006); Sturgeon (2008); Walley (1991). Bradley (2019) is a good source of further references. Imprecise probabilism is closely related to what we might call interval probabilism (Kyburg, 1961; Kyburg Jr & Teng, 2001). In interval probabilism, precise probabilities are replaced by intervals of probabilities. On imprecise probabilism, instead, precise probabilities are replaced by sets of probabilities. This makes imprecise probabilism more general since the representor set need not be an interval.

> three of them have bias .4 and one of them has bias .6. You pick a coin at random and plan to toss it. You should be three times more confident that the probability of getting heads is .4. rather than .6.

The first situation can be easily represented by imprecise probabilism. The representor would contain two probability measures, one that assigns .4. and the other that assigns .6 to the hypothesis 'this coin lands heads'. However imprecise probabilism cannot represent the second situation. Since the probability measures in the set are all compatible with the agent's evidence, no probability measure can be assigned a greater (higher-order) probability than any other.[4]

These examples show that imprecise probabilism is not expressive enough to model the scenario of uneven bias. Defenders of imprecise probabilism might concede this point but prefer their account for reasons of simplicity. They could also point out that imprecise probabilism models scenarios that precise probabilism cannot model, for example, a state of complete lack of evidence. In this respect, imprecise probabilism outperforms precise probabilism in expressive power, and also retains theoretical simplicity. But this is not quite true as imprecise probabilism suffers from several shortcomings that do not affect precise probabilism.

The first shortcoming we discuss has not received extensive discussion in the literature. Recall that, for imprecise probabilism, an agent's state of uncertainty is represented by those probability measures that are *compatible* with the agent's evidence. How should the notion of compatibility be understood here? Perhaps we can think of compatibility as the fact that the agent's evidence is consistent with the probability measure in question. But mere consistency wouldn't get the agent very far in excluding probability measures, as too many probability measures are consistent with most observations and data. Admittedly, there will be clear-cut cases: if you see the outcome of a coin toss to be heads, you reject the measure with $\mathrm{P}(H) = 0$, and similarly for tails. Another class of cases might arise while randomly drawing objects from a finite set where the objective chances are known. But clear-cut cases aside, what else? Data will often be consistent with almost any probability measure.[5]

A second, related problem for imprecise probabilism is known as belief inertia. Precise probabilism offers an elegant model of learning from evidence: Bayesian updating. Imprecise probabilism, at least *prima facie*, offers an equally elegant model of learning from evidence,

[4] Other scenarios can be constructed in which imprecise probabilism fails to capture distinctive intuitions about evidence and uncertainty; see, for example, (Rinard, 2013). Suppose you know of two urns, GREEN and MYSTERY. You are certain GREEN contains only green marbles, but have no information about MYSTERY. A marble will be drawn at random from each. You should be certain that the marble drawn from GREEN will be green (G), and you should be more confident about this than about the proposition that the marble from MYSTERY will be green (M). For each $r \in [0, 1]$ your representor will contain a P with $\mathrm{P}(M) = r$. But then, it also contains one with $\mathrm{P}(M) = 1$. This means that it is not the case that for any probability measure P in the representor set, $\mathrm{P}(G) > \mathrm{P}(M)$, that is, it is not the case that a rational agent should be more confident of G than of M. This is counter-intuitive.

[5] Probability measures can be inconsistent with evidential constraints that agents believe to be true. Mathematically, non-trivial evidential constraints are easy to model (Bradley, 2012). They can take the form, for example, of the *evidence of chances* $\{\mathrm{P}(X) = x\}$ or $\mathrm{P}(X) \in [x, y]$, or *structural constraints* such as "X and Y are independent" or "X is more likely than Y." These constraints are something that an agent can come to accept outright, but only if offered such information by an expert whom the agent completely defers. Besides these idealized cases, it is unclear how an agent could come to accept such structural constraints upon observation. There will usually be some degree of uncertainty about the acceptability of these constraints.

richer and more nuanced. It is a natural extension of the classical Bayesian approach that uses precise probabilities. When faced with new evidence E between time t_0 and t_1, the representor set should be updated point-wise, running the standard Bayesian updating on each probability measure in the representor:

$$\mathbb{P}_{t_1} = \{\mathrm{P}_{t_1} | \exists\, \mathrm{P}_{t_0} \in \mathbb{P}_{t_0} \ \forall\, H \ \left[\mathrm{P}_{t_1}(H) = \mathrm{P}_{t_0}(H|E)\right]\}.$$

The hope is that, if we start with a range of probabilities that is not extremely wide, point-wise learning will behave appropriately. For instance, if we start with a prior probability of *heads* equal to .4 or .6, then those measures should be updated to something closer to .5 once we learn that a given coin has already been tossed ten times with the observed number of heads equal 5 (call this evidence E). This would mean that if the initial range of values was $[.4, .6]$ the posterior range of values should be narrower.

Unfortunately, this narrowing of the range of values becomes impossible whenever the starting point is a complete lack of knowledge, as imprecise probabilism runs into the problem of belief inertia (Levi, 1980). This problem arises in situations in which no amount of evidence could lead the agent to change their belief state, according to a given modeling strategy. Consider a situation in which you start tossing a coin knowing nothing about its bias. The range of possibilities is $[0, 1]$. After a few tosses, if you observed at least one tail and one head, you can exclude the measures assigning 0 or 1 to *heads*. But what else have you learned? If you are to update your representor set point-wise, you will end up with the same representor set. For any sequence of outcomes that you can obtain and any probability value in $[0, 1]$, there will exist a probability measure (conditional on the outcomes) that assigns that probability to *heads*. Consequently, the edges of your resulting interval will remain the same. In the end, it is not clear how you are supposed to learn anything if you start from complete ignorance.[6]

Some downplay the problem of belief inertia. After all, if you start with knowing truly nothing, then it is right to conclude that you will never learn anything. Joyce (2010) writes:

> You cannot learn anything in cases of pronounced ignorance simply because a prerequisite for learning is to have prior views about how potential data should alter your beliefs. (p. 291) (Joyce, 2010)

The upshot is that uniform priors should not be used and that imprecise probabilism gives the right results when the priors are non-vacuous. Moss (2020) arrives at the same conclusion by drawing the following parallelism. If contingent propositions should not be assigned probabilities of 0 or 1 whenever these extreme values are unrevisable, then by the same token the uniform interval [0, 1] should not be used for imprecise probabilities whenever it is impervious to revision.[7] However, as we will see in the next section, uniform priors are not necessarily

[6] Here is another example of belief inertia by Rinard (2013). Suppose all the marbles in the urn are green (H_1), or exactly one-tenth of the marbles are green (H_2). Your initial credence about these hypotheses is completely uncertain, the interval $[0, 1]$. Next, you learn that a marble drawn at random from the urn is green (E). After conditioning on this evidence, you end up with the same spread of values for H_1 that you had before learning E. This result holds no matter how many green marbles are drawn. This is counterintuitive: if you keep drawing green marbles, the hypothesis that all marbles are green should become more probable.

[7] Another strategy is to say that, in a state of complete ignorance, a special updating rule should be deployed.

unrevisable and can be a starting point for learning. This suggests that the problem lies with imprecise probabilism, not with uniform priors as such.

Finally, even setting aside belief inertia, imprecise probabilism faces a third major problem that does not arise for precise probabilism. Workable scoring rules exist for measuring the accuracy of a precise probability measure, but it is hard to define workable scoring rules for imprecise probabilities. In the precise case, scoring rules measure the distance between a rational agent's probability measure and the actual value. The Brier score is the most common scoring rule for precise probabilities.[8] A requirement of scoring rules is *propriety*: any rational agent will expect their probability measure to be more accurate than any other. After all, if an agent thought a different probability measure was more accurate, they should switch to it. More specifically, let $I(p, w)$ be an inaccuracy score of a probability distribution p relative to the true state w. The score $I(p, w)$ is strictly proper if, for any other probability distribution q different from p, the following holds:

$$\sum_{w \in W} p(w)I(p, w) < \sum_{w} q(w)I(p, w).$$

That is, the expected inaccuracy of p from the perspective of p should always be smaller than the expected inaccuracy of p from the perspective of another distribution q. To calculate the expected accuracy of p, first you need to calculate its inaccuracy $I(p, w)$ at every possible true state w, and then factor in the probabilities $p(w)$ of the true states according to p. Well-known results demonstrate the strict propriety of the Brier score for precise probabilities.[9]

Can similar results be established for imprecise probabilities? The answer is likely to be negative. Several hurdles exist. To start, a plausible scoring rule for imprecise probabilities is not easy to define. Suppose a forecaster assigns the [.8, .9] probability interval to the outcome that it would rain tomorrow, where the true state is 'rain'. Would the wider [.6, .99] interval be more accurate since its .99 upper bound is closer to the true state? Intuitively, the wider interval should not be more accurate. If it were, the trivial interval [0, 1] would always be more accurate than any other interval since either of its edges are closer to whatever the true state turns out to be. To remedy this problem, an inaccuracy score for imprecise probabilities should depend both on the closeness to the true state and the size of the interval. If, for example, the inaccuracy score were to increase as the size of the interval increases, this would block the result that the [0, 1] interval is always the least inaccurate. Still, even if a well-behaved score for imprecise probabilities can be found, a further problem arises in defining its expected inaccu-

Elkin (2017) suggests the rule of *credal set replacement* that recommends that upon receiving evidence the agent should drop measures rendered implausible, and add all non-extreme plausible probability measures. This, however, is tricky. One needs a separate account of what makes a distribution plausible from a principled account of why one should use a separate special update rule when starting with complete ignorance.

[8] The Brier score is defined as the squared distance between the true state and the probability forecast, or formally, $(p(x) - V(x, w))^2$, where $p(x)$ is the probability forecast and $V(x, w)$ determines if a proposition obtains at w ($V(x, w) = 1$) or not ($V(x, w) = 0$). If, for example, the proposition 'rain' obtains at w, the forecast 'rain with .9 probability' would be more accurate at w than the forecast 'rain with .8 probability'. If, on the other hand, the proposition 'not rain' obtained at w, the latter forecast would be more accurate.

[9] Besides propriety, other common requirements (which the Brier score also satisfies) are: the score $I(p, w)$ should be a function of the probability distribution p and the true state w (extensionality); and the score should be a continuous function around p (continuity).

racy. Let $I([p_-, p_+], w)$ be an inaccuracy score for the interval $[p_-, p_+]$. What is its expected inaccuracy from the perspective of, say, the interval $[p_-, p_+]$ itself? There is no straightforward answer to this question because $I([p_-, p_+], w)$ cannot be multiplied by $[p_-, p_+]$, in the way in which $I(p, w)$ can be multiplied by $p(w)$. Perhaps the expected inaccuracy of $[p_-, p_+]$ can be evaluated from the perspective of the precise probabilities at its edges, either p_- or p_+.[10] The problem is that, if the expected inaccuracy of $[p_-, p_+]$ is evaluated from the perspective of the precise probabilities at the edges, finding a proper inaccuracy score that is also continuous turns out to be mathematically impossible (Seidenfeld, Schervish, & Kadane, 2012).[11]

4 Higher-order probabilism

Let us take stock. Imprecise probabilism is more expressive than precise probabilism. It can model the difference between a state in which there is no evidence about a proposition (or its negation) and a state in which the evidence for and against a proposition is in equipoise. However, imprecise probabilism has its own expressive limitations, for example, it cannot model the case of uneven bias. In addition, imprecise probabilism faces difficulties that do not affect precise probabilism: the notion of compatibility between a probability measure and the evidence is too permissive; belief inertia makes it impossible for a rational agent to learn via Bayesian updating; and no proper scoring rules exist for imprecise probabilism. In this section and the next, we show that higher-order probabilism overcomes the expressive limitations of imprecise probabilism without falling prey to any such difficulties.

Proponents of imprecise probabilism already hinted at the need to rely on higher-order probabilities. For instance, Bradley compares the measures in a representor to committee members, each voting on a particular issue, say the true chance or bias of a coin. As they acquire more evidence, the committee members will often converge on a chance hypothesis.

> …the committee members are "bunching up". Whatever measure you put over the set of probability functions—whatever "second order probability" you use—the "mass" of this measure gets more and more concentrated around the true chance hypothesis. (Bradley, 2012, p. 157)

But such bunching up cannot be modeled by imprecise probabilism alone: a probability distribution over chance hypotheses is needed.[12] That one should use higher-order probabilities has also been suggested by critics of imprecise probabilism. Carr (2020) argues that sometimes evidence requires uncertainty about what credences to have. Carr, however, does not

[10] So the expected inaccuracy would equal $\sum_{w \in W} p_-(w) I([p_-, p_+], w)$ or $\sum_{w \in W} p_+(w) I([p_-, p_+], w)$.

[11] Proper scoring rules are often used to formulate accuracy-based arguments for precise probabilism. These arguments show (roughly) that, if your precise measure follows the axioms of probability theory, no other non-probabilistic measure is going to be more accurate than yours whatever the facts are. So, without proper scoring rules for imprecise probabilities, the prospects for an accuracy-based argument for imprecise probabilism look dim (Campbell-Moore, 2020; Mayo-Wilson & Wheeler, 2016). Moreover, as shown by Schoenfield (2017), if an accuracy measure satisfies certain plausible formal constraints, it will never strictly recommend an imprecise stance, as for any imprecise stance there will be a precise one with at least the same accuracy.

[12] In a similar vein, Joyce (2005), in a paper defending imprecise probabilism, explicates the notion of weight of evidence using a probability distribution over chance hypotheses. Oddly, representor sets play no central role in Joyce's account of the weight of evidence.

articulate this suggestion more fully; does not develop it formally; and does not explain how her approach would fare against the difficulties affecting precise and imprecise probabilism. We now set out to do precisely that.

The central idea of higher-order probabilism is this: a rational agent's uncertainty cannot be mapped onto a one-dimensional scale such as the real line. Uncertainty is best modeled by the shape of a probability distribution. In some straightforward cases of narrow and symmetric distributions, we can get away with using point probabilities, but such approximations will fail to be useful in more complex cases. So, special cases aside, a rational agent's state of uncertainty (or credal stance) towards a proposition x is not represented by a single probability value $p(x)$ between 0 and 1, but by a probability density $f(p(x))$, where the first-order probability of x is the parameter in question and is treated as a random variable. Crucially, this representation is general. While the examples used so far may not indicate this, the first-order probability of x is not restricted to chance hypotheses or the bias of a coin. The probability density $f(p(x))$ assigns a second-order probability (density) to all possible first-order probabilities $p(x)$.[13]

How should these second-order probabilities be understood? It is helpful to think of higher-order probabilism as a generalization of imprecise probabilism. Imprecisers already admit that some probability measures are compatible and others incompatible with the agent's evidence at some point. Compatibility is a coarse notion; it is an all-or-nothing affair. But, as seen earlier, evidence can hardly exclude a probability measure in a definitive manner except in clear-cut cases. Just as it is often a matter of degrees whether the evidence supports a proposition, the compatibility between evidence and a probability measure can itself be a matter of degrees. In this picture, the evidence justifies different values of first-order probability to various degrees. So, second-order probabilities express the extent to which the first-order probabilities are supported by the evidence.

This higher-order approach at the technical level is by no means novel. Bayesian probabilistic programming languages embrace the well-known idea that parameters can be stacked and depend on each other (Bingham et al., 2021). But, while the technical machinery has been around for a while, it has not been deployed by philosophers to model a rational agent's uncertainty or credal state. Because of its greater expressive power, higher-order probabilism can represent uncertainty in a more fine-grained manner, as illustrated in Figure 1. In particular, the uneven coin scenario in which the two biases of the coin are not equally likely—which imprecise probabilism cannot model—can be easily modeled within high-order probabilism by assigning different probabilities to the two biases.

An agent's uncertainty could—perhaps, should—sometimes be represented by a single probability value. Higher-order probabilism does not prohibit that. For example, there may well be cases in which an agent's uncertainty is aptly represented by the expectation.[14] But this need not always be the case. If the probability distribution is not sufficiently concentrated around a single value, a one-point summary will fail to do justice to the nuances of the agent's credal

[13] For computational ease, we will use a higher-order density that is discretized and ranges over 1000 first-order probabilities.

[14] The expectation is usually defined as $\int_0^1 x f(x)\, dx$. In the context of our approach here, x is the first-order probability of a given proposition and f is the density representing the agent's uncertainty about x.

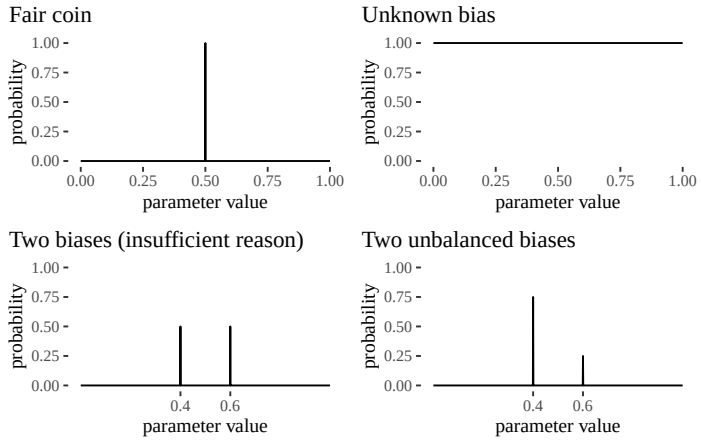

Figure 1: Examples of higher-order distributions for a few scenarios problematic for both precise and imprecise probabilism.

state.[15] For example, consider again the scenario in which the agent knows that the bias of the coin is either .4 or .6 but the former is three times more likely. Representing the agent's credal state with the expectation $.75 \times .4 + .25 \times .6 = .45$ would fail to capture the agent's different epistemic attitudes towards the two biases. The agent believes the two biases have different probabilities, and is also certain the bias is *not* .45.

Besides its greater expressive power in modeling uncertainty, higher-order probabilism does not fall prey to belief inertia. Consider a situation in which you have no idea about the bias of a coin. You start with a uniform distribution over $[0, 1]$ as your prior. Observing any non-zero number of heads will exclude 0 and observing any non-zero number of tails will exclude 1 from the basis of the posterior. The posterior distribution will become more centered as the observations come in. This result is a straightforward application of Bayesian updating. Instead of plugging sharp probability values into the formula for Bayes's theorem, the factors to be multiplied in the theorem will be probability densities (or ratios of densities as needed). Figure 2 illustrates—starting with a uniform prior distribution—how the posterior (beta) distribution changes after successive observations.[16]

The impossibility of defining proper scoring rules was another weakness of imprecise probabilism. This is a significant shortcoming, especially because proper scores do exist for precise probabilism. Fortunately, one can show that there exist proper scoring rules for higher-order

[15] This approach lines up with common practice in Bayesian statistics, where the primary role of uncertainty representation is assigned to the whole distribution. Summaries such as the mean, mode standard deviation, mean absolute deviation or highest posterior density intervals are only succinct ways of representing uncertainty.

[16] Assuming independence and constant probability for all the observations, learning is modeled the Bayesian way. You start with some prior density p over the parameter values. If you start with a complete lack of information, p should be uniform. Then, you observe the data D which is the number of successes s in a certain number of observations n. For each particular possible value θ of the parameter, the probability of D conditional on θ follows the binomial distribution. The probability of D is obtained by integration. That is:

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \\ &= \frac{\theta^s(1-\theta)^{(n-s)}p(\theta)}{\int(\theta')^s(1-\theta')^{(n-s)}p(\theta')\ d\theta'}. \end{aligned}$$

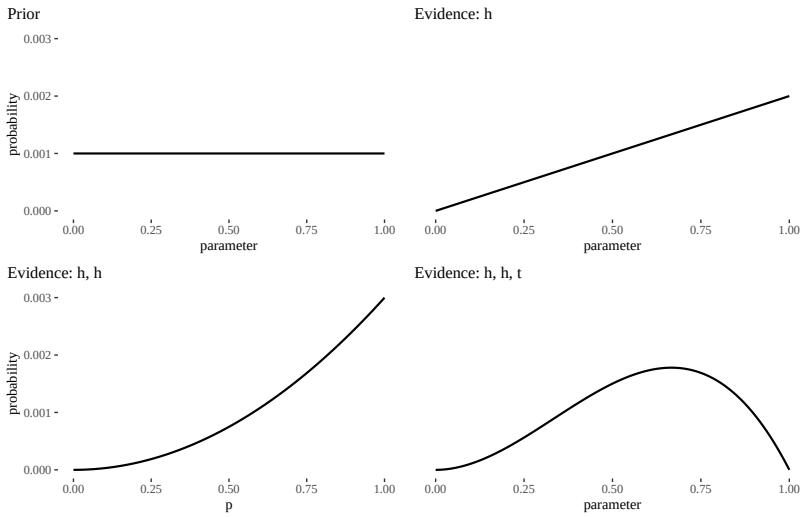

Figure 2: As observations of heads, heads and tails come in, extreme parameter values drop out of the picture and the posterior is shaped by the evidence.

probabilism. We defend this claim in the next section. The argument, however, will be more formal and can be skipped upon first reading.

5 Higher-order proper scores

Despite the difficulties that plague imprecise probabilism in defining proper scoring rules, here we put forward an intuitively plausible scoring rule for higher-order probabilities that is both proper and continuous. Building on existing work on this topic (Hersbach (2000), Pettigrew (2012), Gneiting & Raftery (2007)), we begin by laying out two common measures of distance between probabilities distributions. We then use these measures to define higher-order inaccuracy scores and argue for their propriety.

As a measure of the distance between two distributions p and q, the Cramer-Von-Mises (CM) measure is a natural starting point. It is defined as follows:

$$D_{\text{CM}}(p, q) = \sum_x |P(x) - Q(x)|^2,$$

where x ranges over all hypotheses under consideration (i.e. the elements of the sample space). The CM measure sums over the square of the differences between $P(x)$ and $Q(x)$ for each value x, where P and Q are the cumulative distributions corresponding to the probability distribution p and q. Looking at cumulative densities is a technical requirement that ensures that all densities are considered on the same scale. As an alternative measure, the Kullback-Leibler (KL) divergence is a common information-theoretic measure of distance between probability distributions. The distance between p and q from the perspective of p is defined as follows:

$$D_{\text{KL}}(p||q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

The KL measure sums over the log of the ratio of $p(x)$ to $q(x)$ for each value x. Note that the KL measure contains a weighing by the distributions p, while the CM measure does not. For ease of computation, these two measures are discretized.[17]

Both measures can be turned into inaccuracy scores provided one of the two distributions plays the role of the higher-order distribution whose accuracy is to be measured and the other distribution tracks the true state of the world. But before moving forward, some notation is needed. Since many of the examples in this paper are about coin tosses and their biases, let $\theta_1, \ldots, \theta_n$ be a finite set of hypotheses about the bias of a coin (this is our discretization of the sample space). Depending on the state of the world, one of these hypotheses will correspond to the true coin bias, call it θ_k. Each θ_k is paired with an omniscient distribution $Ind^k(\cdot)$, such that $Ind^k(\theta_i)$ is 1 if $i = k$ and 0 otherwise. In other words, since the omniscient distribution tracks the true state of the world (i.e. the true bias being θ_k), it will assign 1 to the true chance hypothesis θ_k and 0 to all the others. For simplicity, we will write Ind^k_i instead of $Ind^k(\theta_i)$.

With this notation in place, the inaccuracy of a higher-order probability distribution p if the true state is θ_k can be defined using the CM measure, as follows:

$$I_{\text{CM}}(p, \theta_k) = \sum_{i=1}^{n} |\text{P}(\theta_i) - Ind^k_i|^2$$

Using instead KL divergence between p and Ind^k, the inaccuracy of a higher-order probability distribution p if the true state is θ_k can be defined, as follows:

$$I_{\text{KL}}(p, \theta_k) = \sum_{i=1}^{n} Ind^k_i \log \left(\frac{Ind^k_i}{p(\theta_i)} \right)$$

As shown in the appendix, $I_{\text{KL}}(p, \theta_k)$ boils down to $-\log p(\theta_k)$. If, for example, the true bias of the coin is .6 and the higher-order distribution p assigns .8 to this bias, the higher-order inaccuracy score of p would be $-\log(.8)$.[18]

To check that the inaccuracy scores just defined work as intended, consider a variation of a scenario by Schoenfield (2017). A rational agent is invited to engage in a bet by an opponent who has a representative bag of coins coming from a factory where the distribution of bias among the coins produced, the true generative process, is known. It is a mixture of two normal distributions centered at .3 and at .5, both with a standard deviation of .05. The opponent randomly selects one of the coins in the bag and flips it. The rational agent who knows this set-up may form a number of higher-order credal states in response to this information. Consider three such credal states, out of many options: first, a faithful bimodal distribution centered at

[17] In the continuous case, KL divergence is defined as the differential KL divergence and the CM measure as the area under the squared Euclidean distances between the corresponding cumulative density functions. That is, $D_{CM}(p, q) = \int_0^1 |P(x) - Q(x)|^2 \, dx$. There are no readily computable solutions to this integral, although it can sometimes be evaluated in the closed form (Gneiting & Raftery, 2007, p. 366).

[18] On this approach, two distributions p and p' which assign the same probability to the true coin bias will have the same inaccuracy score even though they might differ in the probabilities they assign to other possible coin biases. So the shape of the distribution does not matter for the inaccuracy score, but it does matter for expected inaccuracy (more on this soon).

.3 and .5; second, a unimodal distribution centered at .4; third, a wide bimodal distribution centered at .2 and .6. The three options are depicted in Figure 3. All of them have expected values at about .4. So, if precise probabilities were the only measure of uncertainty, .4 would be the most natural value to assign to the probability that coin came up, say, heads. The three distributions, however, differ in how they represent higher-order uncertainty, and it seems that the faithful bimodal distribution gives the best representation, a point to which we will return.

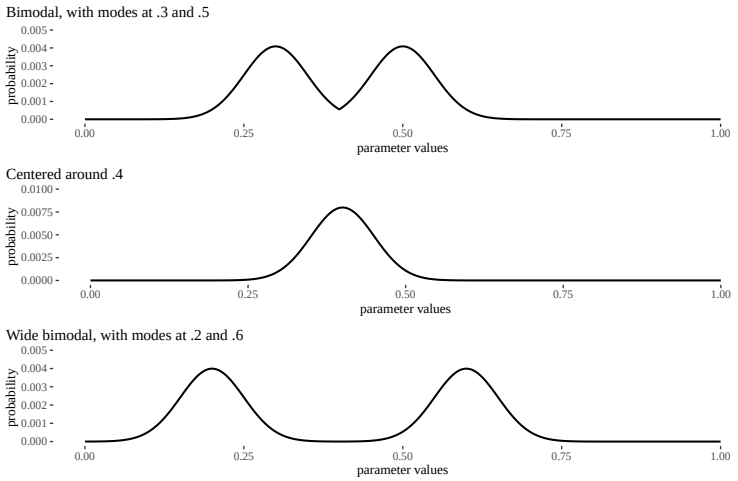

Figure 3: Three distributions in a vague EMS scenario. The distributions are built from normal distributions with standard deviation .05, the bimodal ones are joint in the middle. All of them have expected values $\approx .4$.

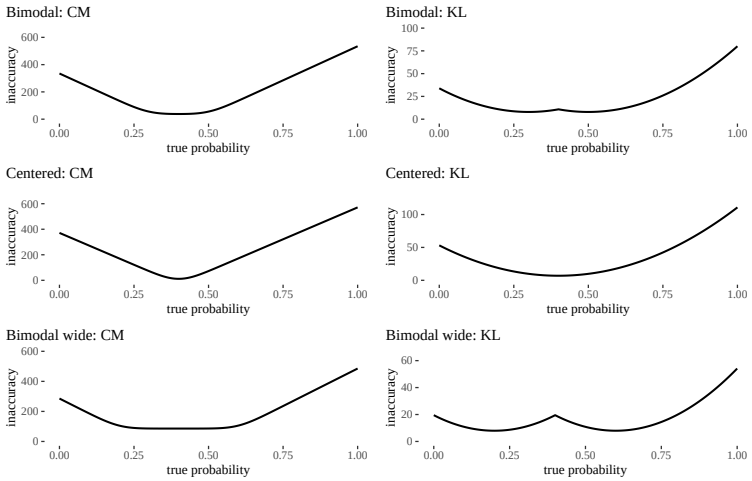

Figure 4: CM and KL divergence based inaccuracies relative to n true chance hypotheses for the three distributions, faithful bimodal, centered unimodal and wide bimodal.

The accuracy scores of these higher-order distributions are in Figure 4. Each point in the graph reflects the accuracy score calculated relative to a possible omniscient distribution corresponding to the values of θ, the true bias of the coin. Both inaccuracy scores behave as intended. The inaccuracy scores are higher at the extremes: if the true coin bias is indeed close to 1 (the coin is weighted to heads) or 0 (the coin is weighted to tails), the three distribu-

tions consider these biases extremely unlikely. However, an important difference transpires between the CR and KL measures. For chance hypotheses between the peaks of the two bimodal distributions, the CR measure remains flat, an artifice of using a squared distance metric. By contrast, the KL-based inaccuracy score jumps slightly for values in between the peaks. This outcome is more intuitive, and a reason to prefer KL-based inaccuracy scores.

To complete our discussion, the final step is argue for the propriety of the scoring rules I_{KL} and I_{CM}. The higher-order score $I(p, \theta_k)$, whether in the KL or CM version, is strictly proper if, for any other probability distribution q different from p, the following holds:

$$\sum_{k=1}^{n} p(\theta_k) I(p, \theta_k) < \sum_{k=1}^{n} q(\theta_k) I(p, \theta_k).$$

That is, the expected inaccuracy of p must be lower when evaluated from the perspective of itself compared to any other distribution q. That the inequality holds is confirmed by simulations in our running example. The expected inaccuracies, in the KL and CM versions, of the three distributions—faithful bimodal, wide bimodal and unimodal—from their own perspective, as well as from the perspective of the other distributions, are in Table 1. The results show that from their own perspective, the distributions see themselves as the least inaccurate.

	CM			KL		
	bimodal	centered	wide bimodal	bimodal	centered	wide bimodal
bimodal	64.670	78.145	88.380	8.577	10.655	11.336
centered	41.657	28.181	85.911	9.239	7.690	15.627
wide bimodal	137.699	171.719	113.989	11.541	19.231	8.689

Table 1: Expected inaccuracies of the three distributions from their own perspective and that of the other distributions. Each row corresponds to a perspective.

To generalize this argument, we prove the strict propriety of the KL-based inaccuracy measure in the appendix. The gist of the proof is this: the expected KL-based inaccuracy of p from the perspective of p itself equals the entropy of p, denoted by $H(p)$, while the expected inaccuracy of p from the perspective of a different distribution q equals the cross-entropy $H(p, q)$. Since $H(p) < H(p, q)$ always holds by Gibb's inequality when $p \neq q$, the KL-based inaccuracy of a distribution from its own perspective will always be the lowest.

A corollary of the propriety of the KL- or CM-based scoring rules is that the faithful bimodal distribution should be preferred over the others. The unimodal distribution, while centering on the expected value, gets the chances wrong, and the wide bimodal has its guesses too close to the true values and too far from the known chances. So the faithful bimodal seems the most evidence-responsive. How can this intuition be captured formally? The expected inaccuracy of each distribution should be measured from the perspective of the true generative process, which we know to be the faithful bimodal centered at .3 and .5. By strict propriety, the expected inaccuracy of the faithful bimodal is the lowest, a good reason to prefer it over the others.[19]

[19] Alternatively, note that the KL-distance or CM-distance between the faithful bimodal distribution and the true

6 Conjunctions

Here is where we are. We have seen that imprecise probabilities model uncertainty better than precise ones. But imprecise probabilities fall short in their own way, for example, when the biases of a coin are not equally likely given the evidence available. Higher-order probabilities are better able to model these more complex scenarios. They also avoid the pitfalls of imprecise probabilities, such as belief inertia and the difficulty of finding proper scoring rules.

One limitation of the discussion so far, however, is that we only looked at assessing probabilities of individual events, say whether a coin would come up heads or tails. But, of course, rational agents may need to assign probabilities to multiple events, for example, the conjunction of two events. Suppose I am holding two coins, and I have information about their respective biases. What is, then, the probability that they both come up, say, heads? In the precise case, the answer is straightforward: assuming independence, it is enough to multiply the individual probabilities. But what happens in the imprecise case? And how to proceed with higher-order probabilities? Once again, we will see that in assessing probabilities for conjunctions of events higher-order probabilities fare better than precise and imprecise ones.

Instead of relying on coin tosses, we will go through a legal example. We selected this example to illustrate how higher-order probabilities can be useful beyond cases of coin tossing. In a murder case, the police recover two items of so-called match evidence: first, hair found at the crime scene matches the defendant's hair; and second, the fur of the defendant's dog matches the fur found in a carpet wrapped around one of the bodies.[20] These two matches constitute evidence against the defendant. The most obvious explanation is that the defendant visited the crime scene and contributed both traces. The alternative explanation is that the matches are a coincidence. Maybe another person visited the scene and happened to have the same hair type and a dog with the same fur type. How likely is that? Trial experts usually provide coincidental match probabilities (also called random match probabilities). They express the likelihood that, by coincidence, a random person (or a random dog) who is not a contributor would still match. If the coincidental match probabilities are low, the two matches would be strong incriminating evidence. If they are not low—the hair type and dog fur type are common—the two matches would be weak incriminating evidence.

It is customary to rely on database frequencies to assess the coincidental match probabilities, for example, by counting how many matches are found in a sample of the human population or the canine population. Suppose the matching hair type occurs 0.0253 times in a reference database, and the matching dog fur type occurs 0.0256 times in a reference database (more on how these numbers are calculated soon). These frequencies give the individual coincidental probabilities. To assess the probability of the two coincidental matches happening jointly, it is enough to multiply the individual probabilities: $0.0253 \times 0.0256 = 6.48 \times 10^{-4}$. Multiplication is allowed on the assumption that the coincidental matches are independent events.[21] The

generative process (which is the faithful bimodal itself), is by definition zero, while it is greater than zero for the other distributions. So, again, the faithful bimodal should be preferred.

[20] The hair evidence and the dog fur evidence are stylized after two items of evidence in the notorious 1981 Wayne Williams case (Deadman, 1984b, 1984a).

[21] The two matches are independent conditional on the hypothesis that the defendant is not a contributor.

resulting joint probability is very small. The two matches, combined, are strong evidence against the defendant, or so it would appear.

This is the story told by the precise probabilist. But this story misses something crucial. As it happens, the coincidental match probability for hair evidence is based on 29 matches found in a sample database of size $1,148$, while the coincidental match probability for the dog evidence is based on finding 2 matches in a smaller database of size 78. The relative frequencies are about .025 in both cases, but the two samples differ in size. The smaller the sample, the greater the uncertainty about the match probabilities. So, for individual pieces of evidence, simply reporting the exact numbers makes it seem as though the evidential value of the hair and fur matches is the same, but actually, it is not.[22] In the aggregate, multiplying the coincidental match probabilities further washes away this difference.

A better approach is available: take into account higher-order uncertainty. Figure 5 (upper part) depicts higher-order probability distributions of different coincidental match probabilities given the sample data—the actual number of matches found in the sample databases. As expected, some coincidental probabilities are more likely than others, and since the sizes of the two databases are different, the distributions have different spreads: the smaller the database the greater the spread, the greater the uncertainty about the coincidental probability. In light of this, Figure 5 (lower part) depicts the probability distribution for the joint coincidental match probabilities associated with both hair and fur evidence. The mathematics here is straightforward: once the higher-order distributions are known, simply multiply them to obtain the higher-order distribution of the joint coincidental match probabilities.

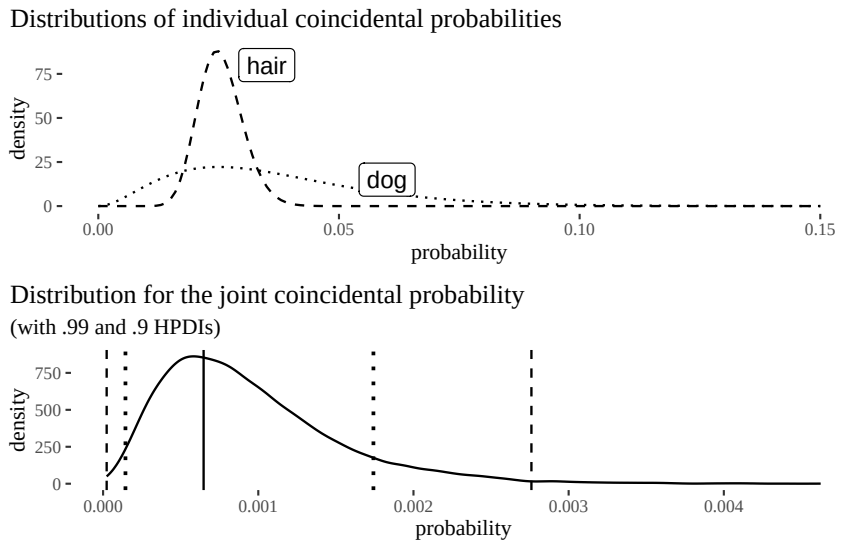

Figure 5: Beta densities for individual items of evidence and the resulting joint density with .99 and .9 highest posterior density intervals, assuming the sample sizes as discussed and independence, with uniform priors.

[22] The probabilities in the Wayne Williams case on which our running example is based were $1/100$ for the dog fur, and $29/1148$ for Wayne Williams' hair. Probabilities have been slightly but not unrealistically modified to be closer to each other in order to make a conceptual point: the same first-order probabilities, even when they sound precise, may come with different degrees of second-order uncertainty.

The precise probabilist might insist that our best assessment of the first-order coincidental match probabilities is still the relative frequency of matches found in the database, whether large or small. All things considered, our best assessment of the match probabilities for both fur and hair evidence should be about .025, based on the relative frequencies 2/78 and 29/1,148. After all, if we were to bet whether a dog or a human picked at random would have the matching fur or hair type, our odds should be .025 no matter the size of the database. This argument has some bite for individual events. In fact, the expected values of the coincidental probabilities for hair and match evidence—based on the higher-order distributions in Figure 5 (upper part)—still end up being about .025. If, as the precise probabilist assumes, first-order probabilities are all we should care about, going higher-order would seem a needless complication.

This line of reasoning, however, breaks down when evaluating conjunctions of events. What should our betting odds be for the proposition that a human and a dog, both picked at random, would have the matching fur and hair type in question? For the precise probabilist, the answer is straightforward: on the assumption of independence, multiply the .025 individual match probabilities and obtain a joint match probability of 6.48×10^{-4}. The higher-order probabilist will proceed differently. In assessing first-order match probabilities, they will retain information about higher-order uncertainty as much as possible. This can done in two steps: first, aggregate the higher-order distributions for the two-match probabilities and obtain a higher-order distribution for the joint match probability (see Figure 5); next, to obtain our best assessment of the first-order joint match probability, take the expected value of this latter distribution. Interestingly, the higher-order probabilist will assign 9.38×10^{-4} to the joint coincidental match probability, a value greater than what the precise probabilist would assign.

So, the higher-order and precise probabilist will disagree about the betting odds for the proposition that a human and a dog, both picked at random, would have the matching fur and hair type in question. The disagreement will become even starker as a larger number of independent items of evidence are evaluated.[23] Who should be trusted? Since the higher-order probabilist takes into account more information—that is, information about the strength of evidence as reflected in sample size—there is good reason to think that the higher-order probabilist should be trusted more than the precise probabilist.

We have not considered yet how imprecise probabilism fares in assessing the probabilities of multiple events in conjunction. Recall that the probability measures in the representor set are those compatible with the evidence. Now, almost any coincidental match probability will be compatible with any sample data. Think by analogy to coin tossing: even a coin that has a .99 bias toward tails could come up heads on every toss. This series of outcomes is unlikely but possible. Similarly, a hair type that has a match probability extremely small could still be found several times in a sample population. So the appropriate interval would be [0,1] for both coincidental match probabilities, and the same for the conjunction. This result would make it

[23] Consider the simple case of independent items of evidence whose individual match probabilities are .025. For three, five and seven items of evidence, the joint match probabilities would be: 1.25×10^{-4}, 3×10^{-7} and 8×10^{-10} (for the precise probabilist); 5.34×10^{-4}, 1.68×10^{-5} and 9.999×10^{-7} (for the higher-order probabilist, based on small databases of size 20); and 1.08×10^{-4}, 2×10^{-7} and 6×10^{-10} (for the higher-order probabilist, based on larger databases of size 20,000).

impossible to learn anything. So this route is a non-starter.

Suppose instead we rely on reasonable ranges of coincidental match probabilities, for example, (.015,.037) (.002, .103), for hair and fur match evidence respectively.[24] As expected, the range is set wider for dog fur match evidence than hair match evidence: the uncertainty about the dog fur match probability is greater since the sample database was smaller. This is a good feature of the interval approach, unavailable to the precise probabilist. But how to assess the joint uncertainty? The most natural strategy is to focus on what happens at the edges of the two intervals. Reasoning with representor members at the edges of the intervals will yield the most extreme probability measure the impreciser is committed to, the worst-case and best-case scenarios. Following this strategy yields a new range for the joint match probabilities: (.00003, .003811).[25] Since relying on ranges for the match probabilities leaves the impression that any value in the interval is just as good as any other, perhaps we should pick the middle value as representative of the interval. But consider again Figure 5 (lower part) which depicts the probability distribution for the joint match probability. This distribution is not symmetric: the most likely value and the bulk of the distribution do not lie in the middle between the edges. So, only relying on the edges—or taking central values as representative—can lead to overestimating or underestimating the probabilities at play.[26]

All in all, precise and imprecise probabilism does not fare well in assessing the probabilities of conjunction of independent events. In the case of individual events, this problem might not be as apparent, but when the probabilities of multiple events are assessed, the divergence between higher-order probabilism and the other versions of probabilism becomes starker. Insisting that all we should care about are first-order probabilities will not work if the values of the first-order probabilities are not assessed in light of all the information available. What precise and imprecise probabilities are ultimately guilty of is neglecting useful information.

One final clarification. While the examples in this section involve match probabilities based on samples of different sizes, the problem we are highlighting is not confined to differences in sample size or match probabilities; it is broader than that. Probabilities can be subject to higher-order uncertainty for other reasons, for example, when they are derived from a probability model that has little support, or when the sample size is unrepresentative. While assessing the probabilities of events in conjunction, these higher-order uncertainties may be compounded. It would be a mistake to ignore them, even if all we cared about were first-order probabilities.

[24] These are 99% credible intervals using uniform priors. A 99% credible interval is the narrowest interval to which the expert thinks the true parameter belongs with probability .99. On credible intervals, see Kruschke (2015).

[25] Redoing the calculations using the upper bounds of the two intervals yields $.037 \times .103 = .003811$. This number is around 5.88 times greater than the original, precise estimate. The calculation for the lower bounds yields $.015 \times .002 = .00003$. This number is around 0.46 times lower than the original estimate.

[26] The calculations for the joint interval assume that because the worst- or best-case probability for one event is x and the worst- or best-case probability for another independent event is y, the worst- or best-case probability for their conjunction is xy. However, this conclusion does not follow if the margin of error (credible interval) is fixed. Just because the probability of an extreme value x for one variable X is .01, and so it is for the value y of another independent variable Y, it does not follow that the probability that those two independent variables take values x and y simultaneously is the same. In general, it is impossible to calculate the credible interval for the joint distribution based solely on the individual credible intervals corresponding to the individual events.

7 Bayesian networks

We looked at simple cases of conjunctions in which the events in question were probabilistically independent. But realistic scenarios are more complex. Think, for example, of two witnesses testifying in a trial about two propositions, say the defendant's whereabouts and the defendant's motive. These two propositions are likely probabilistically dependent. To model these more complex cases, precise probabilists will rely on Bayesian networks, compact representations of probability distributions over several random variables. The graphical part of a Bayesian network consists of nodes and arrows. Arrows between nodes visually represent relationships of probabilistic dependence between different hypotheses and items of evidence, each corresponding to a node (variable) in the network. The numerical part of a Bayesian network describes the strengths of these dependencies via suitable conditional probabilities. Equipped with these input conditional probabilities, the network can run the calculations about the other conditional probabilities of interest. We might be interested, for example, in the probability that the defendant did this-or-that given several items of evidence, while keeping track of dependencies between them.[27] In their standard formulation, Bayesian networks run on precise probabilities but can be extended to handle imprecise and higher-order probabilities.

As an illustration, let us start with a Bayesian network developed by Fenton & Neil (2018). The network in Figure 6 represents the key items of evidence in the infamous British case R. v. Clark (EWCA Crim 54, 2000). Sally Clark, the mother of two sons, witnessed her first son die in 1996 soon after birth. Her second son died in similar circumstances a few years later in 1998. These two consecutive deaths raised suspicion. One hypothesis about the cause of death is that Sally murdered her children. An alternative explanation is that both children died of Sudden Infant Death Syndrome (SIDS). At trial, however, the pediatrician Roy Meadow testified that the probability that a child from a family like the Clark's would die of SIDS was quite low, 1 in 8,543. Assuming probabilistic independence between the two events, the probability of both children dying of SIDS equals the product of the two probabilities, approximately 1 in 73 million, an extremely unlikely event. Based on this low probability and signs of bruising on the bodies, Sally Clark was convicted of murder. The conviction was reversed on appeal thanks to new, exculpatory evidence that was later found.

Much has been written about Sally Clark by philosophers and statisticians. The discussion has often focused on whether Meadow was allowed to assume, as he did, that the two SIDS deaths would be independent events. The assumption of independence delivered the low probability of 1 in 73 million by squaring the figure 1 in 8,543. Another much-discussed point was that, even if it was unlikely that two consecutive SIDS deaths would occur, it does not follow it was likely that Sally murdered her children.[28] A Bayesian network helps to avoid these mistakes. It also helps to view the case holistically. The two consecutive deaths were an important piece of evidence, but other evidence was also important, including signs of bruising and signs of a lethal disease as they were discovered during the appeal process.

[27] The calculations can quickly get out of hand, so software exists to perform the calculations automatically.

[28] One could reason that, since 1 in 73 million is a low probability, the alternative explanation, that Sally murdered her children, should be likely. But that a mother's killing her children is also unlikely.

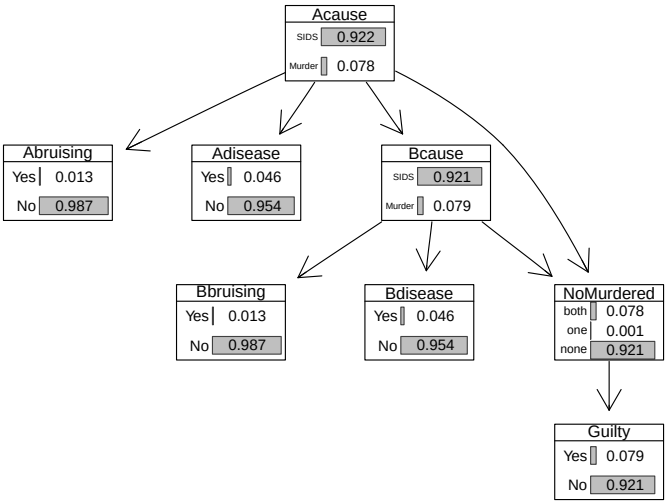

Figure 6: Bayesian network for the Sally Clark case, with marginal prior probabilities.

Unfortunately, Bayesian networks, in their standard formulation, inherit the shortcomings of precise probabilism. The input probabilities should be precise, but it is often unclear where the values come from or whether they are justified. Consider the probability that a death by SIDS would occur. How sure are we that this probability equals 1 in 8,543? Meadow's figure was based on a sample. How big was that sample? How representative? Other input probabilities need to be entered into the network to carry out the calculations, for example, the probability that a mother would kill her children, or the probability that signs of bruising would be found if Sally was trying to murder her children, and so on. As they are based on sample frequencies or expert elicitation, these probabilities will also be uncertain.

The standard response to these concerns is to run a *sensitivity analysis*: a range of plausible values is tested. Say we are interested in the output probability that Sally is guilty. The network is updated by the known facts—the items of evidence—following standard Bayesian conditionalization. The input probabilities in the network are then assigned a range of possible values to see how they impact the output probability of Sally's guilt. Sensitivity analysis is another variant, perhaps more rudimentary, of imprecise probabilism. In fact, Bayesian networks for reasoning with intervals and imprecise probabilities already exist.[29] But, as discussed earlier, imprecise probabilism ignores the shape of the underlying distributions. It does not distinguish between probability measures in terms of their plausibility, even though some will be more plausible than others. Moreover, if the sensitivity analysis is only guided by the values at the edges of the interval, these extremes will often play an undeservedly strong role.

These concerns can be addressed by recourse to higher-order probabilities. In a precise Bayesian network, each node is associated with a probability table filled in with a finite list of numbers (precise probabilities). In an imprecise Bayesian network, each node is associated

[29] One can use uniform sampling with Bayesian networks to approximate the impreciser's commitments (Caprio et al., 2024). Another approach is to rely on probabilistic programs with the restriction that the variables corresponding to probabilities are sampled from uniform distributions corresponding to the representor set. A critical survey of approaches along these lines shows that, in complex reasoning situations, "the imprecision of inferences increases rapidly as new premises are added to an argument" (Kleiter, 1996).

with a table filled in with an interval of numbers. Instead of precise numbers or intervals, the probability tables can be filled in with distributions over the possible first-order probabilities.[30] An example of such a higher-order Bayesian network for the Sally Clark case can be found in Figure 7. This network helps to assess the impact of the items of evidence on the ultimate issue, Sally Clark's guilt. The answer is significantly uncertain even though this might not be apparent by just looking at the first-order probability of guilt (for details, see Figure 8).[31] The upshot is that relying on precise probabilities only can lead to overconfidence.

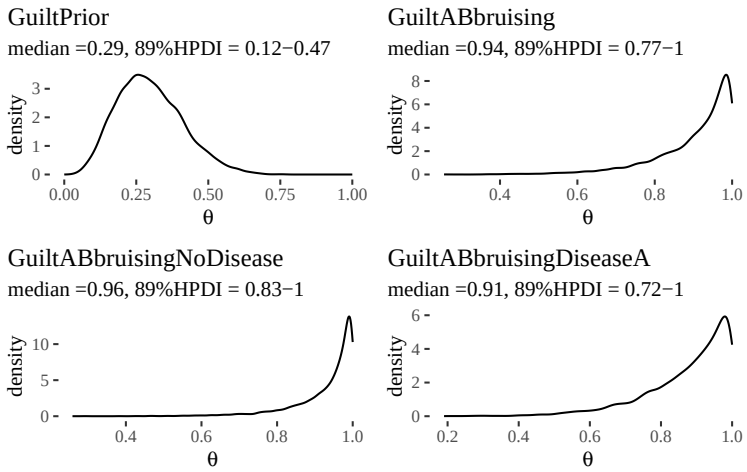

Figure 8: Impact of incoming evidence in the Sally Clark case.

8 Conclusion

We have argued that higher-order probabilism outperforms both precise and imprecise probabilism. It can model scenarios that the other two cannot model, for example, the case of uneven bias. In addition, higher-order probabilism does not fall prey to difficulties peculiar to imprecise probabilism, such as belief inertia and the lack of proper scoring rules. We have also identified a novel set of problems for precise and imprecise probabilism, mostly stemming from the question of how to evaluate, in the aggregate, the probabilities of multiple propositions. Here again, higher-order probabilism fares better.

Some might dislike the idea of going higher-order for several reasons, for example, unnecessary complexity. This is a line taken by Bradley (2019):

> Why are sets of probabilities the right level to stop the regress at? Why not sets of sets? Why not second-order probabilities? Why not single probability functions?

[30] The densities of interests can then be approximated by (1) sampling parameter values from the specified distributions, (2) plugging them into the construction of the BN, and (3) evaluating the probability of interest in that precise BN. The list of the probabilities thus obtained will approximate the density of interest.

[31] The starting point is the prior distribution for the Guilt node (first graph). Next, the network is updated with evidence showing signs of bruising on both children (second graph). Next, the assumption that both children lack signs of potentially lethal disease is added (third graph). Finally, we consider the state of the evidence at the time of the appellate case: signs of bruising existed on both children, but signs of lethal disease were discovered only in the first child. Interestingly, in the strongest case against Sally Clark (third graph), the median of the posterior distribution is above .95, but the uncertainty around that median is still quite wide. (The lower limit of the 89% Highest Posterior Density Intervals (HPDI) is at .83.)

> This is something of a pragmatic choice. The further we allow this regress to continue, the harder it is to deal with these belief-representing objects. So let's not go further than we need (pp. 131-132).

But, given the difficulties of precise and imprecise probabilism, we are not going further than we need in introducing higher-order probabilities. The pragmatic concerns are at best unclear.

We should underscore that, mathematically, we do not propose anything radically new. Concepts from the Bayesian toolkit that can model higher-order uncertainty already exist. We suggest that they have been under-appreciated in formal epistemology and should be more widely used. This is not to say that there is no need for any novel technical work. One concern is the lack of clear semantics for higher-order probabilities. While a more elaborate account is beyond the scope of this paper, the answer should gesture at a modification of the framework of probabilistic frames (Dorst, 2022b, 2022a).[32] Another concern is the lack of an accuracy-based argument in defense of higher-order probabilism. Will an agent who relies on higher-order probabilities always accuracy-dominate one who relies on just first-order probabilities? We leave this as an open question.

Appendix: the strict propriety of I_{KL}

The fact that I_{KL} is strictly proper for second-order probabilities is not very surprising. However, the proof is not usually explicitly given in the existing literature. So we include below the whole chain of reasoning, but we note that some of these results are already common knowledge. Let us start with a definition of concavity.

Definition 1 (concavity)**.** *A function f is convex over an interval (a, b) just in case for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$ we have:*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

A function f is concave just in case:

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2)$$

A function f is strictly concave just in case the equality holds only if either $\lambda = 0$ or $\lambda = 1$.

For us it is important that if a function is twice differentiable on an interval, then it is (strictly) concave just in case its second derivative is non-positive (negative). In particular,

[32] Start with a set of possible worlds W. Suppose you consider a class of probability distributions D, a finite list of atomic sentences $q_1, \ldots, q_2$ corresponding to subsets of W, and a selection of true probability hypotheses C (think of the latter as omniscient distributions, $C \subseteq D$, but in principle this restriction can be dropped if need be). Each possible world $w \in W$ and a proposition $p \subseteq W$ come with their true probability distribution, $C_{w,p} \in D$ corresponding to the true probability of p in w, and the distribution that the expert assigns to p in w, $P_{w,p} \in D$. Then, various propositions involving distributions can be seen as sets of possible worlds, for instance, the proposition that the expert assigns d to p is the set of worlds w such that $P_{w,p} = d$. There is at least one important difference between this approach and that developed by Dorst. His framework is untyped, which allows for an enlightening discussion of the principle of reflection and alternatives to it. In this paper, we prefer to keep this complexity side and use an explicitly typed setup.

as $(\log_2(x))'' = -\frac{1}{x^2 ln(2)}$, $\log_2$ is strictly concave over its domain. (We will work with log base 2, but we could equally well use any other basis.)

Lemma 1 (Jensen's inequality)**.** *If f is concave, and g is any function of a random variable, $\mathbb{E}(f(g(x))) \leq f(\mathbb{E}(g(x)))$. If f is strictly concave, the equality holds only if $g(x) = \mathbb{E}g(x)$, that is, if $g(x)$ is constant everywhere.*

Proof. For the base case consider a two-point mass probability function. Then,

$$p_1 f(g(x_1)) + p_2 f(g(x_2)) \leq f(p_1 g(x_1) + p_2 g(x_2))$$

follows directly from the definition of concavity, if we take $\lambda = p_1$, $(1 - \lambda) = p_2$, and substitute $g(x_1)$ and $g(x + 2)$ for x_1 and x_2.

Now, suppose that $p_1 f(g(x_1)) + p_2 f(g(x_2)) = f(p_1 g(x_1) + p_2 g(x_2))$ and that f is strictly concave. That means either $(p_1 = 1 \wedge p_2 = 0)$, or $(p_1 = 0 \wedge p_2 = 1)$. Then either x always takes value x_1, in the former case, or always takes value x_2, in the latter case. $\mathbb{E}g(x) = p_1 g(x_1) + p_2 g(x_2)$, which equals $g(x_1)$ in the former case and $g(x_2)$ in the latter.

Now suppose Jensen's inequality and the consequence of strict contativity holds for $k - 1$ mass points. Write $p'_i = \frac{p_i}{1-p_k}$ for $i = 1, 2, \ldots, k - 1$. We now reason:

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(g(x_i)) &= p_k f(g(x_k)) + (1 - p_k) \sum_{i=1}^{k-1} p'_i f(g(x_i)) \\
&\leq p_k f(g(x_k)) + (1 - p_k) f\left(\sum_{i=1}^{k=i} p'_i g(x_i) \right) && \text{by the induction hypothesis} \\
&\leq f\left(p_k(g(x_k)) + (1 - p_k) \sum_{i=1}^{k-1} p'_i g(x_i) \right) && \text{by the base case} \\
&= f\left(\sum_{i}^{k} p_i g(x_i) \right)
\end{aligned}$$

Notice also that at the induction hypothesis application stage, we know that the equality holds only if $p_k = 1 \vee p + k = 0$. In the former case $g(x)$ always takes value $x_k = \mathbb{E}g(x)$. In the latter case, p_k can be safely ignored and $\sum_{i=1}^{k} p_i g(x_i) = \sum_{i=1}^{k-1} p' g(x_i)$ and by the induction hypothesis we already know that $\mathbb{E}g(x) = g(x)$. □

In particular, the claim holds if we take $g(x)$ to be $\frac{q(x)}{p(x)}$ (were both p and q are probability mass functions), and f to be $\log_2$. Then, given that A is the support set of p, we have:

$$\sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} \leq \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)}$$

Moreover, the equality holds only if $\frac{q(x)}{p(x)}$ is constant, that is, only if p and q are the same pmfs. Let's use this in the proof of the following lemma.

Lemma 2 (Information inequality)**.** *For two probability mass functions p, q, $D_{\mathrm{KL}}(p, q) \geq 0$ with equality iff $p = q$.*

Proof. Let A be the support set of p, and let q be a probability mass function whose support is B.

$$
\begin{aligned}
-D_{\mathrm{KL}}(p, q) &= -\sum_{x \in A} p(x) \log_2 \frac{p(x)}{q(x)} && \text{(by definition)} \\
&= \sum_{x \in A} p(x) - \left(\log_2 p(x) - \log_2 q(x)\right) \\
&= \sum_{x \in A} p(x) \left(\log_2 q(x) - \log_2 p(x)\right) \\
&= \sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} \\
&\leq \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)} && \text{by Jensen's inequality} \\
&\text{(and the equality holds only if } p = q\text{)} \\
&= \log_2 \sum_{x \in A} q(x) \\
&\leq \log_2 \sum_{x \in B} q(x) \\
&= log(1) = 0
\end{aligned}
$$

□

Lemma 3 (decomposition)**.** $D_{\mathrm{KL}} = H(p, q) - H(p)$.

Proof.

$$
\begin{aligned}
D_{\mathrm{KL}}(p, q) &= \sum_{p_i} \left(\log_2 p_i - \log_2 q_i\right) \\
&= -\sum_{p_i} \left(\log_2 q_i - \log_2 p_i\right) \\
&= -\sum_{p_i} \log_2 q_i - \sum_{p_i} -\log_2 p_i \\
&\underbrace{-\sum_{p_i} \log_2 q_i}_{H(p,q)} \underbrace{-\sum_{p_i} \log_2 p_i}_{H(p)}
\end{aligned}
$$

□

With information inequality this easily entails Gibbs' inequality:

Lemma 4 (Gibbs' inequality)**.** $H(p, q) \geq H(p)$ *with identity only if* $p = q$.

We are done with our theoretical set-up, which is already common knowledge. Now we present our argument for the propriety of I_{KL}. Consider a discretization of the parameter space

$[0, 1]$ into n equally spaced values $\theta_1, \ldots, \theta_n$. For each i the 'true' second-order distribution if the true parameter indeed is θ_i—we'll call it the indicator of θ_i— which is defined by

$$Ind^k(\theta_i) = \begin{cases} 1 & \text{if } \theta_i = \theta_k \\ 0 & \text{otherwise} \end{cases}$$

We will write Ind_i^k instead of $Ind^k(\theta_i)$. Now consider a probability distribution p over this parameter space, assigning probabilities $p_1, \ldots, p_n$ to $\theta_1, \ldots, \theta_n$ respectively. It is to be evaluated in terms of inaccuracy from the perspective of a given 'true' value θ_k. The inaccuracy of p if θ_k is the 'true' value, is the divergence between Ind^k and p.

$$\begin{aligned} I_{KL}(p, \theta_k) &= D_{\text{KL}}(Ind^k || p) \\ &= \sum_{i=1}^{n} Ind_i^k \left(\log_2 Ind_i^k - \log_2 p_i\right) \end{aligned}$$

For $j \neq k$ we have $Ind_j^k = 0$ and so $Ind_j^k \left(\log_2 Ind_j^k - \log_2 p_j\right) = 0$. Therefore:

$$= Ind_k^k \left(\log_2 Ind_k^k - \log_2 p_k\right)$$

Further, $Ind_k^k = 1$ and therefore $\log_2 Ind_k^k = 0$, so we simplify:

$$= -\log_2 p_k$$

Finally, the inaccuracy of a distribution q as expected by p, $EI_{\text{DK}}(p, q)$, is defined as follows:

$$\begin{aligned} EI_{\text{DK}}(p, q) &= \sum k = 1^n p_k I_{\text{DK}}(q, \theta_k) \\ &= \sum_{k=1}^{n} p_k \sum_{i=1}^{n} Ind_i^k \left(\log_2 Ind_i^k - \log_2 q_i\right) \\ &= \sum_{k=1}^{n} p_k Ind_k^k \left(\log_2 Ind_k^k - \log_2 q_k\right) \\ &= \sum_{k=1}^{n} p_k (-\log_2 q_k) \\ &= -\sum_{k=1}^{n} p_k \log_2 q_k = H(p, q) \end{aligned}$$

By contrast, the expected inaccuracy of p from its perspective is defined as follows:

$$EI_{\text{DK}}(p, p) = -\sum k = 1^n p_k \log_2 p_k = H(p)$$

References

Bingham, E., Koppel, J., Lew, A., Ness, R., Tavares, Z., Witty, S., & Zucker, J. (2021). Causal probabilistic programming without tears. *Proceedings of the Third Conference on Probabilistic Programming*.

Bradley, S. (2012). *Scientific uncertainty and decision making* (PhD thesis). London School of Economics; Political Science (University of London).

Bradley, S. (2019). Imprecise Probabilities. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019). https://plato.stanford.edu/archives/spr2019/entries/imprecise-probabilities/; Metaphysics Research Lab, Stanford University.

Campbell-Moore, C. (2020). *Accuracy and imprecise probabilities*.

Caprio, M., Dutta, S., Jang, K. J., Lin, V., Ivanov, R., Sokolsky, O., & Lee, I. (2024). *Credal bayesian deep learning*. Retrieved from https://arxiv.org/abs/2302.09656

Carr, J. R. (2020). Imprecise evidence without imprecise credences. *Philosophical Studies*, *177*(9), 2735–2758. https://doi.org/10.1007/s11098-019-01336-7

Deadman, H. A. (1984a). Fiber evidence and the wayne williams trial (conclusion). *FBI L. Enforcement Bull.*, *53*, 10–19.

Deadman, H. A. (1984b). Fiber evidence and the wayne williams trial (part i). *FBI L. Enforcement Bull.*, *53*, 12–20.

Dorst, K. (2022a). Higher-order evidence. In M. Lasonen-Aarnio & C. Littlejohn (Eds.), *The routledge handbook for the philosophy of evidence*. Routledge.

Dorst, K. (2022b). Higher-order uncertainty. In M. Skipper & A. S. Petersen (Eds.), *Higher-order evidence: New essays*.

Elkin, L. (2017). *Imprecise probability in epistemology* (PhD thesis). Ludwig-Maximilians-Universität; Ludwig-Maximilians-Universität München.

Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with bayesian networks*. Chapman; Hall.

Fraassen, B. C. V. (2006). Vague expectation value loss. *Philosophical Studies*, *127*(3), 483–491. https://doi.org/10.1007/s11098-004-7821-2

Gärdenfors, P., & Sahlin, N.-E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese*, *53*(3), 361–386. https://doi.org/10.1007/bf00486156

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

Hare, C. (2010). Take the sugar. *Analysis*, *70*(2), 237–247.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/https://doi.org/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2

Joyce, J. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, *19*(1), 153–178.

Joyce, J. (2010). A defense of imprecise credences in inference and decision Making1. *Philosophical Perspectives*, *24*(1), 281–323. https://doi.org/10.1111/j.1520-8583.2010.00194.

x

Kaplan, J. (1968). Decision theory and the fact-finding process. *Stanford Law Review*, *20*(6), 1065–1092.

Keynes, J. M. (1921). *A treatise on probability, 1921*. London: Macmillan.

Kleiter, G. D. (1996). Propagating imprecise probabilities in bayesian networks. *Artificial Intelligence*, *88*(1), 143–161.

Kruschke, J. (2015). *Doing bayesian data analysis (second edition)*. Boston: Academic Press.

Kyburg, H. E. (1961). *Probability and the logic of rational belief*. Wesleyan University Press.

Kyburg Jr, H. E., & Teng, C. M. (2001). *Uncertain inference*. Cambridge University Press.

Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy*, *71*(13), 391. https://doi.org/10.2307/2025161

Levi, I. (1980). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT Press.

Mayo-Wilson, C., & Wheeler, G. (2016). Scoring imprecise credences: A mildly immodest proposal. *Philosophy and Phenomenological Research*, *92*(1), 55–78. https://doi.org/10.1111/phpr.12256

Moss, S. (2020). Global constraints on imprecise credences: Solving reflection violations, belief inertia, and other puzzles. *Philosophy and Phenomenological Research*, *103*(3), 620–638. https://doi.org/10.1111/phpr.12703

Pettigrew, R. (2012). *Epistemic utility and norms for credences*.

Rinard, S. (2013). Against radical credal imprecision. *Thought: A Journal of Philosophy*, *2*(1), 157–165. https://doi.org/10.1002/tht3.84

Schoenfield, M. (2017). The accuracy and rationality of imprecise credences. *Noûs*, *51*(4), 667–685. https://doi.org/10.1111/nous.12105

Seidenfeld, T., Schervish, M., & Kadane, J. (2012). Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, *53*, 1248–1261. https://doi.org/10.1016/j.ijar.2012.06.018

Sturgeon, S. (2008). Reason and the grain of belief. *Noûs*, *42*(1), 139–165. Retrieved from http://www.jstor.org/stable/25177157

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman; Hall London.

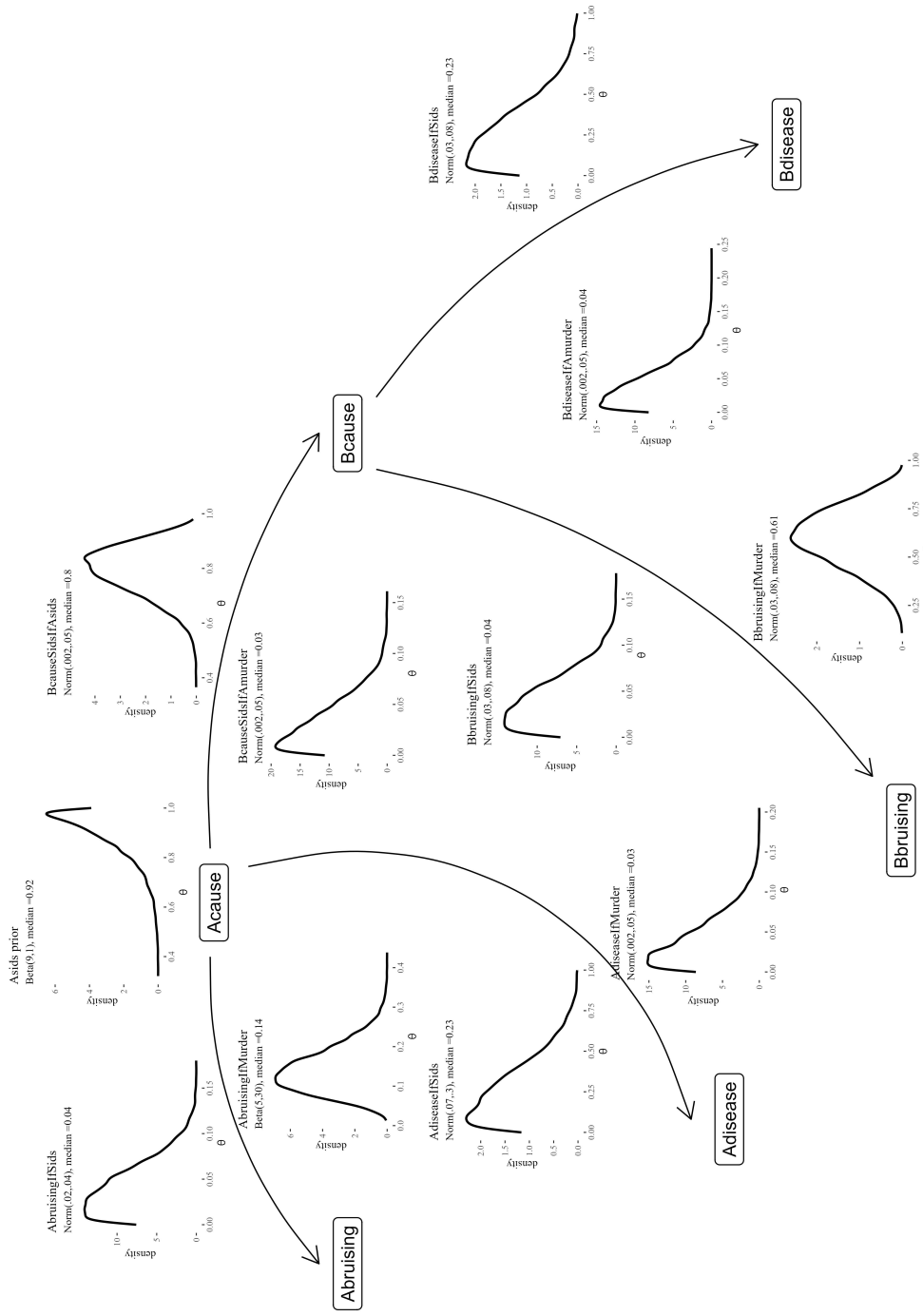

Figure 7: An illustration of a probabilistic program for the Sally Clark case.