Second-order Probability, Accuracy and Weight of Evidence

Rafal Urbaniak and Marcello Di Bello

January 31, 2023

Contents

Abstract.

need to write one when done

1 Introduction

this is what a comment about what will go into book looks like; will be globally supressed when generating latex for the journal paper, don't worry about deleting them.

see M'scomment in boldface

M's comment: This introductory section needs some work. This section is not well-grounded in the literature. Two main things.

FIRST: There is a big debate among forensic scientists about whether intervals shoud be used in the presentation of the value of the evidence. There was a 2016 special issue of Science and Justice on this topic. We should mention it. See, in particular, Stewart Morrison (2016) "Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate" and also, in that special issue, Ommen, Saunders, and Neumann (2016) "An argument against presenting interval quantifications as a surrogate for the value of evidence". Both papers were added to the folder. I am not sure our framing of the debate is faithful to the complexities that are brought up in the forensic science literature.

SECOND: The last paragraph at the end of the section in response to Kadane's remark seems too generic to me. If we are not proposing anything technically new but simply applying existing ideas, we need to cite the technical work we are relying on. Kadane mentioned "Bayesian hierarchical models". I don't know what they are, but I looked up the wikipedia entry on the topic and talked to a statistician. They are a common thing in Bayesian statistics. So we should cite references discussing them and say how we are applying them (if we are just applying them) or say in what way our approach differs from them.

A defendant in a criminal case may face multiple items of incriminating evidence whose strength can at least sometimes be assessed using probabilities. For example, consider a murder case in which the police recover trace evidence that matches the defendant. Hair found at the crime scene matches the defendant's hair (call this evidence hair). In addition, the defendant owns a dog whose fur matches the dog fur found in a carpet wrapped around one of the bodies (call this evidence dog).[1] The two matches suggest that the defendant (and the defendant's dog) must be the source of the crime traces (call this hypothesis source). But how strong is this evidence, really? What are the fact-finders to make of it?

The standard story among legal probabilists goes something like this. To evaluate the strength of the two items of match evidence, we must find the value of the likelihood ratio:

$$\frac{\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \mathsf{source})}{\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \neg \mathsf{source})}$$

For simplicity, the numerator can be equated to one. To fill in the denominator, an expert provides the relevant random match probabilities. Suppose the expert testifies that the probability of a random person's hair matching the reference sample is about 0.0253, and the probability of a random dog's hair matching the reference sample happens to be about the same, 0.0256.[2] Presumably, the two matches are independent lines of evidence. In other words, their random match probabilities must be independent of each other conditional on the source hypothesis. Then, to evaluate the overall impact of the evidence on the source hypothesis, you calculate:

$$\begin{aligned} \mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \neg \mathsf{source}) &= \mathsf{P}(\mathsf{dog} | \neg \mathsf{source}) \times \mathsf{P}(\mathsf{hair} | \neg \mathsf{source}) \\ &= 0.0252613 \times 0.025641 = 6.4772626 \times 10^{-4} \end{aligned}$$

This is a very low number. Two such random matches would be quite a coincidence. Following our advice from Chapter 5, the expert facilitates your understanding of how this low number should be interpreted. They show you how the items of match evidence change the probability of the source hypothesis given a range of possible priors (Figure 1). The posterior of .99 is reached as soon as the prior is higher than 0.061.[3] While perhaps not sufficient for outright belief in the source hypothesis,

[1] 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).

[2] Probabilities have been slightly but not unrealistically modified to be closer to each other in order to make a conceptual point. The original probabilities were 1/100 for the dog fur, and 29/1148 for Wayne Williams' hair. We modified the actual reported probabilities slightly to emphasize the point that we will elaborate further on: the same first-order probabilities, even when they sound precise, may come with different degrees of second-order uncertainty.

[3] These calculations assume that the probability of a match if the suspect and the suspect's dog are the sources is one.

the evidence seems extremely strong: a minor additional piece of evidence could make the case against the defendant overwhelming.

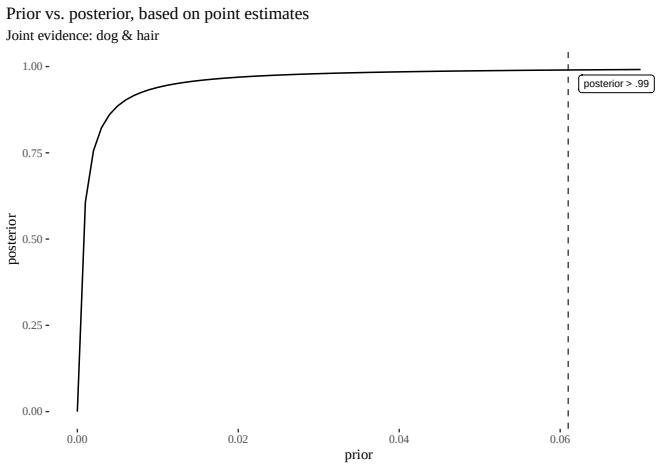

Figure 1: Impact of dog fur and human hair evidence on the prior, point estimates.

Unfortunately, this analysis leaves out something crucial. You reflect on what you have been told and ask the expert: how can you know the random match probabilities with such precision? Shouldn't we also be mindful of the uncertainty that may affect these numbers? The expert agrees, and tells you that in fact the random match probability for the hair evidence is based on 29 matches found in a database of size 1148, while the random match probability for the dog evidence is based on finding two matches in a reference database of size 78.

The expert's answer makes apparent that the precise random match probabilities do not tell the whole story. Perhaps, the information about sample sizes is good enough and now you know how to use the evidence properly.[4] But if you are like most human beings, you can't. What to do, then?

added this bit to draw attention to this aspect of the Taroni debate, to come back to this

You ask the expert for guidance: what are reasonable ranges of the random match probabilities? What are the worst-case and best-case scenarios? The expert responds with 99% credible intervals—specifically, starting with uniform priors, the ranges of the random match probabilities are (.015,.037) for hair evidence and (.002, .103) for fur evidence.[5] With this information, you redo your calculations using the upper bounds of the two intervals: .037 and .103. The rationale for choosing the upper bounds is that these numbers result in random match probabilities that are most favorable to the defendant. Your new calculation yields the following:

$$\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \neg \mathsf{source}) = .037 \times .103 = .003811.$$

This number is around 5.88 times greater than the original estimate. Now the prior probability of the source hypothesis needs to be higher than 0.274 for the posterior probability to be above .99 (Figure 2). So you are no longer convinced that the two items of match evidence are strongly incriminating.

[4] This is what, effectively, CITE TARONI seem to suggest when they insist the fact-finders should be simply given point estimates and information about the study set-up, such as sample size. As will transpire, we disagree.

[5] Roughly, the 99% credible interval is the narrowest interval to which the expert thinks the true parameter belongs with probability .99. For a discussion of what credible intervals are, how they differ from confidence intervals, and why confidence intervals should not be used, see Chapter 3.

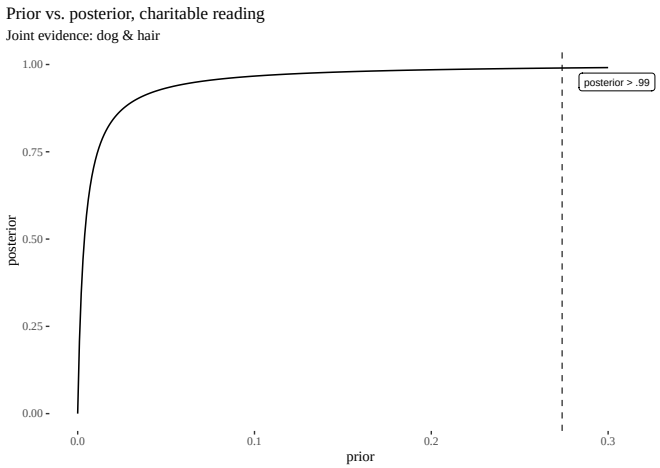

Figure 2: Impact of dog fur and human hair evidence on the prior, charitable reading.

This result is puzzling. Are the two items of match evidence strongly incriminating evidence (as you initially thought) or somewhat weaker (as the new calculation suggests)? For one thing, using precise random match probabilities might be too unfavorable toward the defendant. On the other hand, your new assessment of the evidence based on the upper bounds might be too *favorable* toward them. Is there a middle way that avoids overestimating and underestimating the strength of the evidence?

To see what this middle path looks like, we should reconsider the calculations you just did. You made an important blunder: you assumed that because the worst-case probability for one event is x and the worst-case probability for another independent event is y, the worst-case probability for their conjunction is xy. But this conclusion does not follow if the margin of error (credible interval) is fixed. The intuitive reason is simple: just because the probability of an extreme (or larger absolute) 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. This probability is actually much smaller. The interval presentation instead of doing us good led us into error.

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. We need additional information: the distributions that were used to calculate the intervals for the probabilities of the individual events. In our example, if you additionally knew, for instance, that the expert used beta distributions (as, arguably, they should in this context), you could in principle calculate the 99% credible interval for the joint distribution. It usually will not be the same as whatever the results of multiplication of individual interval edges, and it is unlikely that a human fact-finder would be able to correctly run such calculations in their head even if they knew the functional form of the distributions used. [6] So providing the fact-finder with individual intervals, even if further information about the distributions is provided, might easily mislead.[7]

Can you google to see if there is any such study?

As it turns out, given the reported sample sizes, the 99% credible interval for the probability $\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \neg \mathsf{source})$ is $(0.000023, 0.002760)$.

The upper bound of this interval would then require the prior probability of the source hypothesis to be above .215 for the posterior to be above .99. On this interpretation, the two items of match evidence are still not quite as strong as you initially thought, but stronger than what your second calculation indicated.

the fn was repetitive, compare to fn 5

Still, the interval approach—even the corrected version just outlined—suffers from a more general problem. Working with intervals might be useful if the underlying distributions are fairly symmetrical. But in our case, they might not be. For instance, Figure 3 depicts beta densities for dog fur and human hair, together with sampling-approximated density for the joint evidence. The distribution for the joint evidence is not symmetric. If you were only informed about the edges of the interval, you would be

[6] Also, in principle, in more complex contexts, we need further information about how the items of evidence are related if we cannot take them to be independent.

[7] Investigation of the extent to which the individual interval presentation is misleading would be an interesting psychological study.

oblivious to the fact that the most likely value (and the bulk of the distribution, really) does not simply lie in the middle between the edges. Just because the parameter lies in an interval with some posterior probability, it does not mean that the ranges near the edges of the interval are equally likely—the bulk of the density might very well be closer to one of the edges. Therefore, only relying on the edges can lead one to either overestimate or underestimate the probabilities at play. This also means that—following our advice on how to illustrate the impact of evidence on prior probabilities—a better representation of the dependence of the posterior on the prior should comprise multiple possible sampled lines whose density mirrors the density around the probability of the evidence (Figure 4).

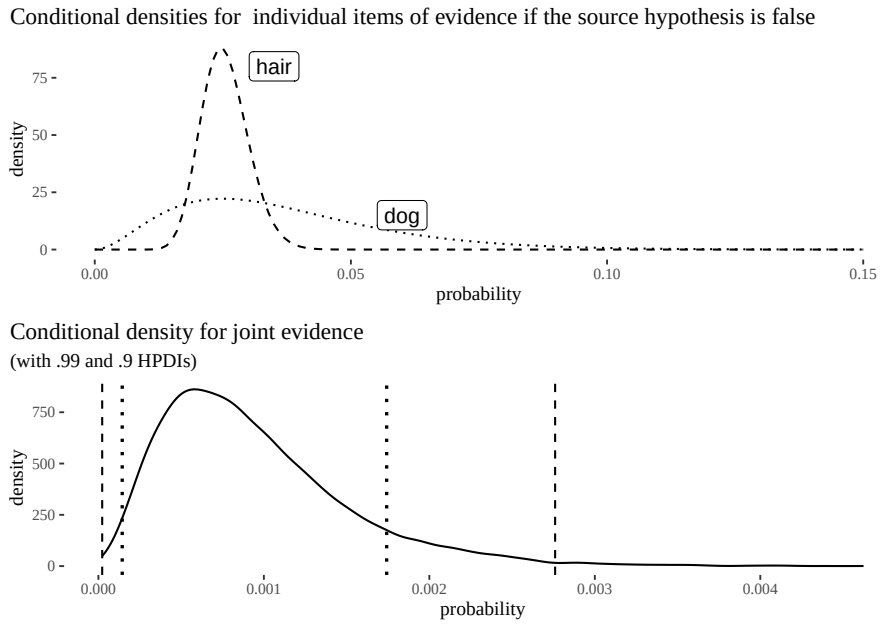

Figure 3: Beta densities for individual items of evidence and the resulting joint density with .99 and .9 highest posterior density intervals, assuming the sample sizes as discussed and independence, with uniform priors.

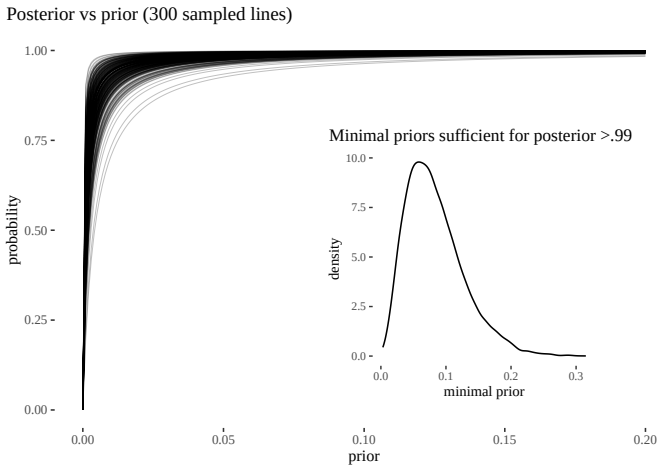

Figure 4: 300 lines illustrating the uncertainty about the dependence of the posterior on the prior given aleatory uncertainty about the evidence, with the distribution of the minimal priors required for the posterior to be above .99.

This, then, is the main claim of this chapter: whenever density estimates for the probabilities of interest are available (and they should be available for match evidence and many other items of scientific

evidence if the reliability of a given type of evidence has been properly studied), those densities should be reported for assessing the strength of the evidence. This approach avoids hiding actual aleatory uncertainties under the carpet. It also allows for a balanced assessment of the evidence, whereas using point estimates or intervals may exaggerate or underestimate the value of the evidence.

In what follows, we expand on this idea in different directions. Section 2 engages with the philosophical debate about precise and imprecise probabilism. We argue that both options are problematic and should be superseded by a higher-order approach to probability whenever possible. Section 3 revisits a recent discussion in the forensic science literature. A prominent view has it that trial experts, even when they use densities, should present only first-order probabilities. We disagree and show that reasons of accuracy maximization sometimes recommend relying on higher-order probabilities. Section 4 turns to some legal applications of higher-order probabilism. We focus on two topics: first, the role of higher-order probabilities and false positive rates in the evaluation of DNA evidence; second, how complex bodies of evidence can be represented by what we call higher-order Bayesian networks.

Before we dive in, one more remark: ost of the time, mathematically, we do not propose anything radically new—we just put together some of the items from the standard Bayesian toolkit. The novelty is rather in our arguing that that these tools are under-appreciated in the legal scholarship and should be properly used to incorporate second-order uncertainties in evidence evaluation and incorporation. Perhaps a minor exception is our explication of the notion of weight, but even here many related notions are available in information theory, and the novelty here is not technical, but rather in the argument that they also are under-appreciated in legal scholarship.

added this par to preemt Kadane's style pickiness

2 Three probabilisms

see M'scomment in boldface

M's comment: this section looks good more or less, but now – given Kadane's remark – there is the question whether the novel proposal – higher-order probabilism – is really new. Is it just an application of hierarchical Bayesian models to discussions in philosophy? In some sense, "probabilism" in epistemology is an application of standard probability theory to questions in epistemology, so perhpas "higher order probabilism" is also an application of existing hiearchical bayesian models to problems in epistemology. If this is right, we need to make these connections clear.

In introduction we sketched three probabilistic approaches that one might take for assessing the value of the evidence presented at trial. The first approach uses precise probabilities; the second uses intervals; the third uses distributions over probabilities. By relying on an example featuring two items of match evidence, we suggested that the third approach is preferable. This section buttresses this claim by providing principled, philosophical reasons in favor of the third approach.

The three approaches we considered correspond (roughly) to three ways in which probabilities can be deployed to model a rational agent's fallible and evidence-based beliefs about the world. The first approach, known in the philosophical literature as precise probabilism, posits that an agent's credal state is modeled by a single, precise probability measure. The second approach, known as imprecise probabilism, replaces precise probabilities by sets of probability measures. The third approach, what we call higher-order probabilism, relies on distributions over parameter values. There are good reasons to abandon precise probabilism and endorse higher-order probabilism. Imprecise probabilism is a step in the right direction, but also suffers from too many difficulties of its own.

2.1 Precise probabilism

Precise probabilism (PP) holds that a rational agent's uncertainty about a hypothesis is to be represented as a single, precise probability measure. This is an elegant and simple theory. But representing our uncertainty about a proposition in terms of a single, precise probability 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 should make the point clear. (For the sake of simplicity, we will use examples featuring coins, but biases of coins can be thought of as random match probabilities in the forensic context.)

> **No evidence v. fair coin** You are about to toss a coin, but have no evidence whatsoever 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 scenario is quite different, but precise probabilities cannot capture this difference.

> **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 unknown bias, you should presumably assign a probability of .5 to both outcomes. After the 10 tosses, you end up again with an estimate of .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. But your precise probability assessment will again be .5.

These examples suggest that precise probabilism is not appropriately responsive to evidence. It ends up assigning the same probability in situations in which one's evidence is quite different: when no evidence is available about the 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). The general problem is, precise probability captures the value around which your uncertainty should be centered, but fails to capture how centered it should be given the evidence.[8]

2.2 Imprecise probabilism

What if we give up the assumption that probability assignments should be precise? Imprecise probabilism (IP) holds that an agent's credal stance towards a hypothesis is to be represented by means of 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. For instance, if an agent knows that the coin is fair, their credal state would be represented by the singleton set $\{\mathsf{P}\}$, where P is a probability measure which 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.[9]

Imprecise probabilism, at least *prima facie*, offers a straightforward picture of learning from evidence, that is a natural extension of the classical Bayesian approach. 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:

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

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 measure 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

[8] Precise probabilism suffers from other difficulties. For example, it has problems with formulating a sensible method of probabilistic opinion aggregation Stewart & Quintana (2018). A seemingly intuitive constraint is that if every member agrees that X and Y are probabilistically independent, the aggregated credence should respect this. But this is hard to achieve if we stick to PP (Dietrich & List, 2016). For instance, a *prima facie* obvious method of linear pooling does not respect this. Consider probabilistic measures p and q such that $p(X) = p(Y) = p(X|Y) = 1/3$ and $q(X) = q(Y) = q(X|Y) = 2/3$. On both measures, taken separately, X and Y are independent. Now take the average, $r = p/2 + q/2$. Then $r(X \cap Y) = 5/18 \neq r(X)r(Y) = 1/4$.

[9] For the development of imprecise probabilism, see Keynes (1921); Levi (1974); Gärdenfors & Sahlin (1982); Kaplan (1968); Joyce (2005); Fraassen (2006); Sturgeon (2008); Walley (1991). S. Bradley (2019) is a good source of further references. Imprecise probabilism shares some similarities with what we might call **interval probabilism** (Kyburg, 1961; Kyburg Jr & Teng, 2001). On 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 probabilities of a proposition in the representor set do not have to form a closed interval. As we have already noted, intervals do not contain probabilistic information sufficient to guide reasoning with multiple items of evidence. So we focus on IP, which is the more promising approach.

more narrow. But even this seemingly straightforward piece of reasoning is hard to model without using densities. For to calculate $\mathsf{P}(\mathsf{heads}|E)$ we need to calculate $\mathsf{P}(E|\mathsf{heads})\mathsf{P}(\mathsf{heads})$ and divide it by $\mathsf{P}(E) = \mathsf{P}(E|\mathsf{heads})\mathsf{P}(\mathsf{heads}) + \mathsf{P}(E) = \mathsf{P}(E|\neg\mathsf{heads})\mathsf{P}(\neg\mathsf{heads})$. The tricky part is obtaining the conditional probabilities $\mathsf{P}(E|\mathsf{heads})$ and $\mathsf{P}(E|\neg\mathsf{heads})$ in a principled manner without explicitly going second-order, estimating the parameter value and using beta distributions.

The situation is even more difficult if we start with complete lack of knowledge, as imprecise probabilism runs into the problem of **belief inertia** (Levi, 1980). Say 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 heads, 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. 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.[10]

Some downplay the problem of belief inertia. They insist that vacuous priors should not be used and that imprecise probabilism gives the right results when the priors are non-vacuous. After all, if you started with knowing truly nothing, then perhaps it is right to conclude that you will never learn anything. Another strategy is to say that, in a state of complete ignorance, a special updating rule should be deployed.[11] But no matter what we think about belief inertia, other problems plague imprecise probabilism. Two more problems are particularly pressing.

One problem is that imprecise probabilism fails to capture intuitions we have about evidence and uncertainty in a number of scenarios. Consider this example:

> **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 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'. But imprecise probabilism cannot represent the second situation, at least not without moving to higher-order probabilities or assigning probabilities to chance hypotheses, in which case it is no longer clear whether the object-level imprecision performs any valuable task.[12]

Second, besides descriptive inadequacy, an even deeper, foundational problem exists for imprecise probabilism. This problem arises when we attempt to measure the accuracy of a representor set of probability measures. Workable *scoring rules* exists for measuring the accuracy of a single, precise credence function, such as the Brier score. These rules measure the distance between one's credence function (or probability measure) and the actual value. A requirement of scoring rules is that they be *proper*: any agent will score their own credence function to be more accurate than every other credence function. After all, if an agent thought a different credence was more accurate, they should switch to it. Proper scoring rules are then used to formulate accuracy-based arguments for precise probabilism. These arguments show (roughly) that, if your precise credence follows the axioms of probability theory, no other credence is going to be more accurate than yours whatever the facts are. Can the same be done for imprecise probabilism? It seems not. Impossibility theorems demonstrate

[10] Here's another example from Rinard (2013). Either all the marbles in the urn are green (H_1), or exactly one tenth of the marbles are green (H_2). Your initial credence $[0, 1]$ in each. Then you learn that a marble drawn at random from the urn is green (E). After conditionalizing each function in your representor on this evidence, you end up with the the same spread of values for H_1 that you had before learning E, and no matter how many marbles are sampled from the urn and found to be green.

[11] 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 or not, as well as a principled account of why one should use a separate special update rule when starting with complete ignorance.

[12] 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). In line with how lack of information is to be represented on IP, for each $r \in [0, 1]$ your representor contains a P with $\mathsf{P}(M) = r$. But then, it also contains one with $\mathsf{P}(M) = 1$. This means that it is not the case that for any probability measure P in your representor, $\mathsf{P}(G) > \mathsf{P}(M)$, that is, it is not the case that RA is more confident of G than of M. This is highly counter-intuitive.

that no proper scoring rules are available for representor sets. So, as many have noted, the prospects for an accuracy-based argument for imprecise probabilism look dim (Campbell-Moore, 2020; Mayo-Wilson & Wheeler, 2016; Schoenfield, 2017; Seidenfeld, Schervish, & Kadane, 2012). 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.

2.3 Higher-order probabilism

There is, however, a view in the neighborhood that fares better: a second-order perspective. In fact, some of the comments by the proponents of imprecise probabilism tend to go in this direction. For instance, Seamus Bradley compares the measures in a representor to committee members, each voting on a particular issue, say the true bias of a coin. As they acquire more evidence, the committee members will often converge on a specific chance hypothesis. He writes (S. Bradley, 2012, p. 157):

> ... 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'.

Note, however, that such bunching up cannot be modeled by imprecise probabilism. Joyce (2005), in a paper defending imprecise probabilism, in fact uses a density over chance hypotheses to account for the notion of evidential weight. The idea that one should use higher-order probabilities has also been suggested by critics of imprecise probabilism. For example, Carr (2020) argues that sometimes evidence requires uncertainty about what credences to have. Carr, however, does not articulate this suggestion more fully, does not develop it formally, and does not explain how her approach would fare against the difficulties affecting precise ad imprecise probabilism.

The key idea of the higher-order approach we propose is that uncertainty is not a single-dimensional thing to be mapped on a single one-dimensional scale such as a real line. It is the whole shape of the whole distribution over parameter values that should be taken under consideration.[13] From this perspective, when an agent is asked about their credal stance towards X, they can refuse to summarize it in terms of a point value $\mathsf{P}(X)$. They can instead express their credal stance in terms of a probability (density) distribution f_x treating $\mathsf{P}(X)$ as a random variable. To be sure, an agent's credal state toward X could sometimes be usefully represented by the expectation

$$\int_0^1 x f(x)\, dx$$

as the precise, object-level credence in X, where f is the probability density over possible object-level probability values. But this need not always be the case. If the probability density f is not sufficiently concentrated around a single value, a one-point summary might fail to do justice to the nuances of the agent's credal state.[14] 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 $\mathsf{P}(X) = .75 \times .4 + .25 \times .6 = .45$ would be inadequate as it would fail to capture the agent's belief that the two biases are uneven.

The higher-order approach can easily model all the challenging scenarios we discussed so far in the manner illustrated in Figure 5. In particular, the 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.

Besides its flexibility in modelling 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. So you start with a uniform density over $[0, 1]$ as your prior. By using binomial probabilities as likelihoods, 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 around

[13] Bradley admits this much (S. Bradley, 2012, p. 90), and so does Konek (Konek, 2013, p. 59). For instance, Konek disagrees with: (1) X is more probable than Y just in case $p(X) > p(Y)$, (2) D positively supports H if $p_D(H) > p(H)$, or (3) A is preferable to B just in case the expected utility of A w.r.t. p is larger than that of B.

[14] 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 for representing the uncertainty of a given scenario. Whether the expectation should be used in betting behavior is a separate problem. Here we focus on epistemic issues.

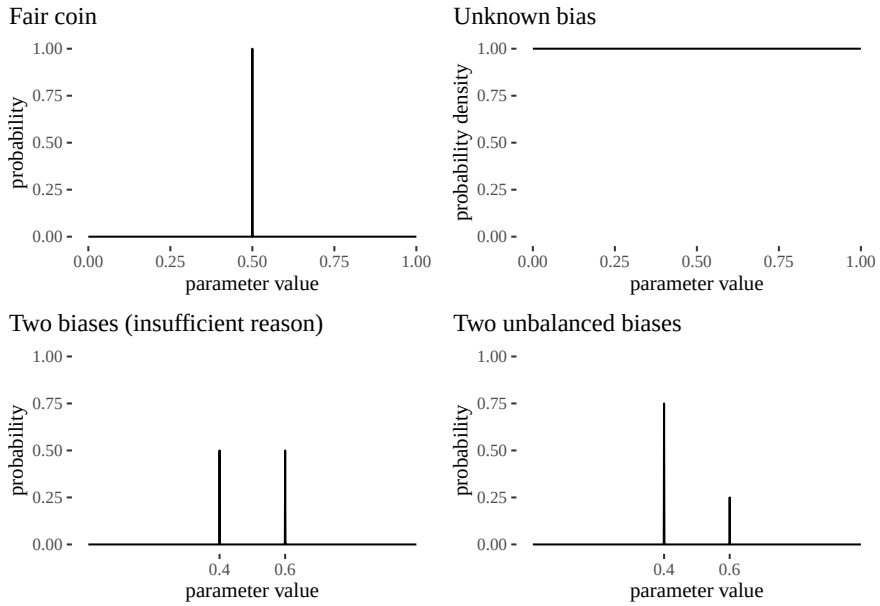

Figure 5: Examples of higher-order distributions for scenarios brought up in the literature.

the parameter estimate as the observations come in. Figure 6 shows—starting with a uniform prior distribution— how the posterior distribution changes after successive observations of heads, heads again, and then tails.[15]

A further advantage of high-order probabilism over imprecise probabilism is that the prospects for accuracy-based arguments are not foreclosed. This is a significant shortcoming of imprecise probabilism, especially because such arguments exist for precise probabilism. One can show that there exist proper scoring rules for higher-order probabilism. These rules can then be used to formulate accuracy-based arguments. Another interesting feature of the framework is that the point made by Schoenfield against imprecise probabilism does not apply: there are cases in which accuracy considerations recommend an imprecise stance (that is, a multi-modal distribution) over a precise one (Urbaniak, 2022 manuscript).

All in all, higher-order probabilism outperforms both precise and imprecise probabilism, at the descriptive as well as the normative level. From a descriptive standpoint, higher-order probabilism can easily model a variety of scenarios that cannot be adequately modeled by the other versions of probabilism. From a normative standpoint, accuracy maximization may sometimes recommend that a rational agent represent their credal state with a distribution over probability values rather than a precise probability measure (more on this in the next section).

3 Objections

This section addresses a number of conceptual difficulties that may arise in using higher-order probabilities, with focus on those brought up by prominent legal evidence scholars. In discussing these conceptual issues, we will formulate an accuracy-based argument that higher-order probabilities are

[15] More generally, learning about frequencies, assuming independence and constant probability for all the observations, is modeled the Bayes way. You start with some prior density p over the parameter values. If you start with 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:

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

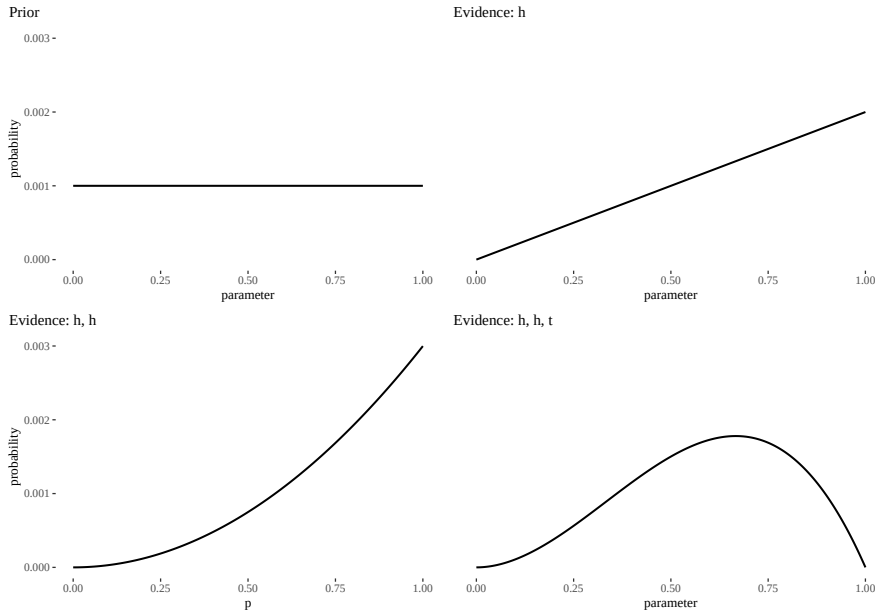

Figure 6: As observations of heads, heads and tails come in, extreme parameter values drop out of the picture and the posterior is shaped by the evidence.

preferable to precise probabilities.

3.1 The Taroni-Sjerps debate

Our treatment will be centered around a discussion initiated by Taroni, Bozza, Biedermann, & Aitken (2015), who argue extensively that trial experts should avoid report higher-order densities, and should only report point estimates. Their point of departure is a reflection on match evidence.

Say an expert reports at trial that the sample from the crime scene matches the defendant. The significance of this match should be evaluated in light of the population frequency θ of the matching profile. This frequency, however, cannot be known for sure and must instead be estimated.

The expert will estimate the true parameter θ by means of a probability distribution $p(\theta)$ over its possible values. For example, if the observations are realizations of independent and identically distributed Bernoulli trails given θ, the expert's uncertainty about θ can be modeled as $\mathtt{beta}(\alpha + s + 1, \beta + n - s)$, where s is the number of observed successes, n the number of observations in the database (1 is added to the first shape parameter to include the match with the suspect), and α and β reflect the expert's priors.

Nothing so far should be controversial. However, the question arises of how the expert should report their own uncertainty about θ, especially in the light of the usual practice of reporting likelihood ratios.

To fix the notation, let the prosecution hypothesis H_p be that the suspect is the source of the trace, and the defense hypothesis H_d that another person, unrelated to the suspect, is the source. For simplicity, assume that if H_p holds, the laboratory will surely report a match M, so that $\mathsf{P}(M|H_p) = 1$. The likelihood ratio, then, reduces to $1/\mathsf{P}(M|H_d)$—but given that θ was estimated using density over its possible values, it is not obvious how a single value $\mathsf{P}(M|H_d)$ is to be obtained and whether its use in the reporting does not hide the uncertainty involved in the estimation of θ under the carpet.

Taroni et al. (2015) claim that the point estimate for the match evidence given the defense hypothesis should be calculated as follows:

$$\begin{aligned}\mathsf{P}(M|H_d) &= \int_\theta \mathsf{P}(M|\theta)\mathsf{P}(\theta)\ d\theta \\ &= \int_\theta \theta\mathsf{P}(\theta)\ d\theta\end{aligned}$$

In case of a DNA match, they recommend that the expert report the expected value of the beta dis-

tribution, which reduces to $^{\alpha+s+1}/_{\alpha+\beta+n+1}$. They claim that this number satisfactorily expresses the posterior uncertainty about θ. For them, it is this probability alone that should be used in the denominator in the calculation and reporting of the likelihood ratio.

Sjerps et al. (2015) disagree. In reporting a single value, the expert would refrain from providing the fact-finders with relevant information that can make a difference in the proper evaluation of the evidence. There is a difference between (a) an expert who is certain θ is .1; (b) an expert whose best estimate of θ is .1 based on thousands of observations; and (c) an expert whose best estimate of θ is again .1 but based on only ten observations.

These three scenarios mirror scenarios we discussed earlier: (a) the bias of a coin is known for sure; (b) the bias is estimated on the basis of a large number of tosses; and (c) the bias is estimated using a small set of observations. As our critique of precise probabilism makes clear, a simple point estimate (or precise probability) would fail to capture the differences among the three scenarios. This concern might be slightly mitigated by the fact that Taroni et al. (2015) admits that the expert, besides providing a point estimate, should also informally explain how the estimate was arrived at. They grant that this additional information can be helpful so long as the recipients are instructed on "the nature of probability, the importance of an understanding of it and its proper use in dealing with uncertainty" [p. 16]. But why stop at an informal presentation? It is unclear why the fact-finders should be deprived of quantifiable information about the aleatory uncertainty of the parameter of interest and only be given an informal description of what the expert did, along with some remarks about the nature of probability. It is wildly optimistic to assume that an informal description of how the point estimate has been arrived at is enough to secure a proper assessment of the evidence. We hope to have convinced the reader already in the introduction that informal treatment and bare intuitions are not good enough even when it comes to the evaluation of the impact of a rather simple combination of two items of evidence if all the fact-finder has to go by is point estimates and and informal description of how the estimates have been obtained.

Somewhat surprisingly, most of the concerns raised by Taroni et al. (2015) are philosophical. They argue that if probabilities express an agent's epistemic attitude towards a proposition probabilities are not states of nature, but states of mind associated with individuals. They think this claim has two consequences. First, it makes no sense sense to talk about second-order uncertainty about subjective probabilities, as there is no "underlying state of the nature" to estimate. Second, if these subjective probabilities can be elicited by examining an agent's betting preferences, a proper elicitation will lead to a single number.[16]

In response to the philosophical argument, Dahlman & Nordgaard (2022) have also emphasized that the distinction is not so clear-cut. They argue that, if a probability assessment is a subjective attitude that is elicited via a betting preference, a probability assessment is itself a state of nature, "the formation of a betting preference by a certain person at a certain time" [p. 15]. While we will have something to say about the philosophical dimension of this debate, let us first develop a less philosophically involved argument for the position taken by Sjerps et al. (2015).

3.2 An accuracy-based argument

see M'scomment in boldface

M's comment: This accuracy-based argument is evocative and intriguing, but what does it show really? What is the significance of using PMF based on a point estimate versus a posterior predictive PMF? Does this correspond to something that is done in court? How? The more interesting question is whether using higher-order probabilities reduces errors, say the rate of false convictions or false acquittals. Does it? If so, how?

With this argument, we hope to break the stalemate in the debate by proving an argument to which both parties should be receptive. It is an accuracy-based argument in favor of using higher-order probabilities—roughly, it says, if you discard relevant information that you already have contained in the densities resulting from the estimation and rely on point estimates only, your predictions about the world will be less accurate in a very precise and quantifiable sense.

First, let us go over a particular example. Suppose we randomly draw a true population frequency

[16] They write: "Clearly, one can adjust the measure of belief of success in the reference gamble in such a way that one will be indifferent with respect to the truth of the event about which one needs to give one's probability. This understanding is fundamental, as it implies that probability is given by a single number. It may be hard to define, but that does not mean that probability does not exist in an individual's mind. One cannot logically have two different numbers because they would reflect different measures of belief." (Taroni et al., 2015, p. 7)

from the uniform distribution. In our particular case, we obtained 0.632. Then, we randomly draw a sample size as a natural number between 10 and 20. In our particular case, it is 16. Next, we simulate an experiment in which we draw that number of observations from the true distribution. We observe 8 successes and use this number to calculate the point estimate of the parameter, which is 0.5.

What is the probability mass function (PMF) for all possible outcomes of an observation of the same size? Two PMF are initially relevant: first, the true probability mass based on the true parameter; second, the probability mass function based on the point estimate which is binomial around the point estimate. This latter PMF, however, does not take into account the uncertainty about the point estimate. To take this uncertainty seriously, continuing our example, we take a sample distribution of size 16 of possible parameter values from the posterior $\mathsf{beta}(1 + \mathsf{successes}, 1 + \mathsf{samplesize} - \mathsf{successes})$ distribution (we assume uniform prior for the sake of an example). Then, we use this sample of parameter values to simulate observations, one simulation for each parameter value in the sample. This simulation yields the so-called *posterior predictive distribution* (or posterior predictive PMF), which instead of a point estimate, propagates the uncertainty about the parameter value into the predictions about the outcomes of possible observations. Finally, we take simulated frequencies as our estimates of probabilities. This distribution is more honest about uncertainty and wider than the one obtained using the point estimate. The three PMFs are displayed in Figure 7.

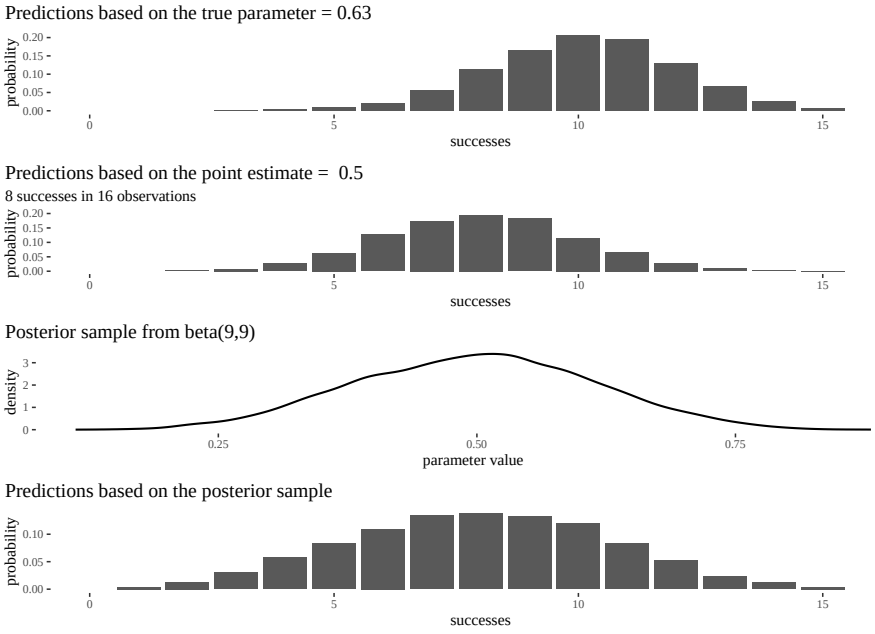

Figure 7: Real probability mass, probability mass calculated using a point estimate, sampling distribution from the posterior, and the posterior predictive distribution based on this sampling distribution.

The PMF based on a point estimate is further off from the real PMF than the posterior predictive distribution. For instance, if we ask about the probability of the outcome being at least 9 successes, the true answer is 0.7984, the point estimate PMF tells us it is 0.4056, while the posterior predictive distribution gives a somewhat better guess at 0.4277. A similar thing happens when we ask about the probability of the outcome being at most 9 successes. The true answer is 0.3681, the point-estimate-based answer is 0.778, while the posterior predictive distribution yields 0.7051. More generally, we can use an information-theoretic measure, Kullback-Leibler divergence, to quantify how far the point-estimate PMF and the posterior predictive PMF are from the true PMF.[17]

[17] A bit of explanation of this divergence measure. Suppose we are dealing with a variable X with n distinct possible discrete states $x_1, \ldots, x_n$ and consider two probability mass functions p and q which express uncertainty about the true value of X so that, say, on p, $\mathsf{P}(X = x_i) = p_i$. First, the uncertainty of a given distribution p, its *entropy*, is given by the sum of the logarithms of surprise $1/p_i$ for all the possible values, $H(p) = \sum x_i \log \frac{1}{p_i} = -\sum p_i \log p_i$. Next, suppose events arise according to p, but we predict them using q. The *cross-entropy* is then $\mathsf{H}(p, q) = \sum p_i \log(q_i)$. This value is going to be higher than the entropy of p if q is different from it. Think of it as the uncertainty involved in using q to predict events that

In our particular case, the former distance is 0.7905638 and the latter is 0.5681121. The posterior predictive distribution is information-theoretically closer to the true distribution.

This was just one example, but the phenomenon generalizes. We repeat the simulation 1000 times, each time with a new true parameter, a new sample size, and a new sample. Every time the three PMFs are constructed using the methods we described and their KL divergence from the true distribution is calculated. Figure 8 displays the empirical distribution of the results of such a simulation. A positive value indicates that the distribution based on the point-estimate was further from the true PMF than the posterior predictive distribution based on the same observed sample. Notably, the mean difference is 0.865, the median difference is 0.044, and the distribution is asymmetrical, as there are multiple cases of large differences favoring posterior predictive distributions over point-based predictions. All in all, accuracy-wise, point-estimate-based PMFs are systematically worse than the posterior predictive distribution.

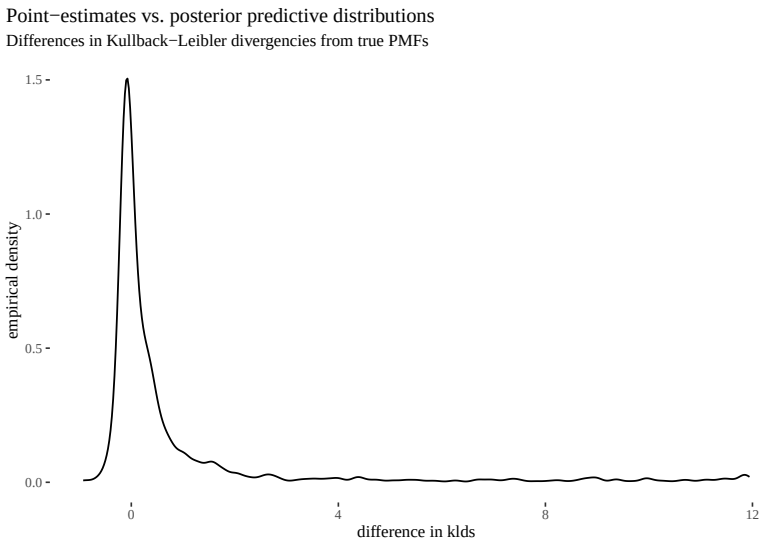

Figure 8: Differences in Kullback-Leibler divergencies from the true distributions, comparing the distributions obtained using point estimates and posterior predictive distributions. Positive values indicate the point-estimate-based PMF was further from the true distribution than the posterior predictive distribution.

3.3 Conceptual issues

M's comment in boldface

M's comments: this subsection looks muddled to me. It goes around in circles. I cannot follow the argument. What is the key point made in this subection? The key point seems to be this: just like we compare the plausibility/probability of propositions like "defendant was at crime scene" and "defedant was not at crime scenbe", we compare the plausibility/probability of propositions like "random match probability is .0001" and "random match probability of .0002". The second comparison requires higher-order probabilities. Is this the point? But Taroni is saying that we

arise according to p. Third, *Kullback-Leibler* divergence is the additional entropy introduced by using q instead of p itself, that is, the difference between cross-entropy and entropy:

$$\begin{aligned}
\mathsf{DKL}(p,q) &= H(p,q) - H(p) \\
&= -\sum p_i \log q_i - \left(-\sum p_i \log p_i\right) \\
&= -\sum p_i \left(\log q_i - \log p_i\right) \\
&= \sum p_i \left(\log p_i - \log q_i\right) \\
&= \sum p_i \log\left(\frac{p_i}{q_i}\right)
\end{aligned}$$

As it turns out, KL divergence is also the expected difference in log probabilities. In particular, if $p = q$ we get $DKL(p,p) = \sum p_i(\log p_i - \log p_i) = 0$, which works out as it intuitively should be.

can average over all possible random match probabilities and get a point estimate. What do we say in response to that?

Accuracy considerations aside, we will now engage with the more conceptual points. Taroni et al. (2015) argue that since first-order probabilities capture your uncertainty about a proposition of interest, second-order probabilities are supposed to capture your uncertainty about how uncertain you are, and that "estimating" your first-order uncertainties is unnecessary. They think that you can simply figure out your fair odds in a suitable bet on the proposition in question, and the fair odds track your unique, first-order uncertainty without any uncertainty about it. But this point can be questioned. For one thing, the betting interpretation of probability is not uncontroversial.[18] Even assuming the betting interpretation, there seems to be nothing wrong in saying that sometimes we are uncertain about what we think the fair bets are.[19] But admittedly, this answer while undermines the betting argument for the sufficiency of point estimates, does not cast much light on what the appropriate relatively uncontroversial interpretation of higher-order uncertainty should be.

add ref to Williamson

M's: I don't understand the argument here. What is a "relatively uncontroversial interpretation"?

Think again about an expert who gathers information about the allelic frequency f of DNA matches in an available database, and starts with a defensible beta prior with parameters α, β. Say the expert observes s matches in a database of size n. So the population relative frequency the experts is estimating should follow the $\mathsf{beta}(\alpha+s+1, \beta+n-s)$ distribution. So far, nothing controversial happens—the expert is estimating the relevant population frequency.

But subjective uncertainty that is to be reported by the expert, Taroni et al. (2015) complain, is not about the frequency, but about their attitude towards a proposition (supposedly expressing a "state of nature")—and, they insist, it makes no sense for an agent to attach uncertainty to their own uncertainty about a proposition.

Assuming the conditions are pristine (the expert has no modeling uncertainty, rules out laboratory errors, and so on), the beta distribution can be used to pretty directly inform the expert's subjective uncertainty. But uncertainty about what? The (estimated) population frequency, for instance, can underlie a probability assignment to the proposition *a match is observed if another person, unrelated to the suspect, is the source of the trace*. Admittedly, if only this proposition is being considered, it is yet not clear what second-order uncertainties would be uncertainties about. But the expert also considers a continuum of propositions, each of the form *the true population frequency is* θ for each $\theta \in [0, 1]$. A density over θ models the comparative plausibility that the expert assigns to such propositions in light of the evidence.[20] So if one were worried that there were no propositions that the expert could be "second-order" uncertain about, there actually are plenty. In particular, if θ is a population frequency, gauging which density captures the extent to which the evidence justifies various estimates of that frequency is the same as gauging the comparative plausibility of the corresponding propositions about possible population frequencies.[21]

M: who was worried that there were no second-order propositions? What argument in the literature is this a response to? I cannot follow.

More generally, in many contexts, evidence justifies first-order probability assignments (population frequency estimates) to various degrees. Suppose there is no evidence about the bias of a coin. Then, each first-order uncertainty about it would be equally (un)-justified. (If you like to think in terms of bets, the evidence would give no reason to prefer any particular odds as fair.) If, instead, we know the coin is fair, the evidence clearly selects one preferred value, .5. (Again, if you like the betting metaphor, 1:1 would be the unique recommended betting odds.) But often the evidence is stronger than the former case and weaker than the latter case. Consider, for example, propositions about population frequencies in light of the results of observations. In such circumstances, the evidence justifies different values of first-order uncertainty to various degrees, and densities simply capture the extent to which different first-order uncertainties are supported by the evidence.

We conclude this section by examining two additional points raised by Taroni et al. (2015). The first—which we already alluded to earlier—is that first-order probabilities are not "states of nature" and so cannot be estimated. It is unclear why the authors insist that only states of nature can be estimated. Mathematicians use approximate methods to estimate answers to fairly abstract questions,

[18] See textbooks in formal epistemology (D. Bradley, 2015; Titelbaum, 2020).

[19] On a related note, the introspective axioms in epistemic logic—that is, if an agent knows (or doesn't know) p, they also know that they (don't) know p—are by no means uncontroversial. See, for example, Williamson 2000 (chapter 5)'s argument against the KK principle of positive introspection.

[20] Moreover, and normalization allows them to calculate their subjective probabilities for θ belonging to various sub-intervals of $[0, 1]$.

[21] Perhaps, this should no longer be called "estimation", but the the connection with estimation is strong enough to justify this terminology. In the end, this is a verbal discussion that we will not get into.

not obviously related to "states of nature", whatever these are. So, estimation should make sense whenever there are some objective answers that we can approximate to a greater or lesser extent. If there is some objectivity to what the ideal evidence would support, or to the extent to which the actual evidence supports various competing hypotheses, we can be more or less wrong about such things, and so it is not implausible to say that there is a clear sense in which we can estimate them.[22]

added back the claim about not worrying about BMI, as some readers might be sensitive to anyone bringing BMI up

Second, Taroni et al. (2015) argue that once we allow second-order probability, we run into the threat of infinite regress. But do we? Surely, they would agree that one can be uncertain about a statistical model. But this can be the case even if this model spits out a point estimate rather than a density. If you think the possibility of putting uncertainty on top of propositions about possible values of a first-order parameter leaves us in an epistemically hopeless situation, you might have hard time explaining why your point estimation is in a better situation. After all, if asking further questions about probabilities up the hierarchy is always justified, we can keep asking about the probability of a point-estimate-spitting model, the probability of that probability, and so on.

Perhaps the problem at issue is just one of complexity. Admittedly, second-order estimation is more complex than relying on point estimates. But we hope to have convinced the reader this complexity is worth the effort. What about more complex models going third-order? If a workable approach can accomplish that—and the additional complexity pays off—we are all for going third-order. The fact that more complex models can always be built hardly lead us into a vicious infinite regress. Rather, it is an indication that our models of uncertainty can—in principle—always be improved.

4 Legal Applications

Add carpet evidence in the Wayne Williams case

Our discussion so far has been mostly theoretical. We made a case that higher-order probabilism outperforms precise probabilism on both descriptive and normative grounds. We also staved off a number of conceptual difficulties with going higher-order. It is time to extend our discussion from the introduction to a further illustration of how higher-order probabilism can be of service in evaluating evidence at trial. We present here two examples.

4.1 False Positives in DNA Identification

One important topic is that of errors in the process of DNA match evidence evaluation. As already known, the probability of a false positive caused by contamination, laboratory or evidence collection or storage error has serious impact on the value of DNA match evidence. As Thompson, Taroni, & Aitken (2003) have shown, the probability of false positives, even when seemingly low, has a non-negligible impact:

> If, as commentators have suggested, the rate of false positives is between 1 in 100 and 1 in 1000, or even less, then one might argue that the jury can safely rule out the prospect that the reported match in their case is due to error and can proceed to consider the probability of a coincidental match . . . this argument is fallacious and profoundly misleading . . . the probability that a reported match occurred due to error in a particular case can be much higher, or lower, than the false positive probability.

We are particularly interested in the passing remark that the rate of false positives is between 1 in 100 and 1 in 1000. This difference is not negligible. The simplest option would be to use the upper bound of the [0.001, 0.01] interval. This choice would be the most favorable toward the defendant. But, as already noted in the introduction, doing so would lead to an overly conservative evaluation of the evidence. It is much preferable to have a sensible distribution to work with.

To fix ideas, the posterior probability of the source hypothesis (S) conditional on the match evidence (E) is, with some idealization, as follows:

[22] Taroni et al. (2015) make the same point for likelihood ratios. They argue that there is no "meaningful state of nature equivalent for the likelihood ratio in its entirety, as it is given by a ratio of two conditional probabilities?" But if it is meaningful to estimate two conditional probabilities (that is, frequencies in the population), or to compare the relative plausibility of various propositions about them in terms of density, it is equally meaningful to estimate any function of the numbers involved. Otherwise it would also be meaningless to try to estimate the body mass index (BMI) of an average 21 years old male student in the USA just because BMI is a ratio of other quantities. There are reasons not to care about BMI, but it not being a state of nature because it is a function of other values is not one of them.

$$
\begin{aligned}
\mathsf{P}(S|E) &= \frac{\mathsf{P}(E|S)\mathsf{P}(S)}{\mathsf{P}(E)} \\
&= \frac{\overbrace{\mathsf{P}(E|S)}^{1}\,\mathsf{P}(S)}{\underbrace{\mathsf{P}(E|S)}_{1}\,\mathsf{P}(S) + \underbrace{\mathsf{P}(E|RM)}_{1}\,\mathsf{P}(RM) + \underbrace{\mathsf{P}(E|FP)}_{1}\,\mathsf{P}(FP)} \\
&= \frac{\mathsf{P}(S)}{\mathsf{P}(S) + \mathsf{P}(RM) + \mathsf{P}(FP)}
\end{aligned}
$$

For simplicity, the false negative rate is assumed to be zero, or in other words, $\mathsf{P}(E|S) = 1$. The other assumption is that the evidence could come about if: (1) the source hypothesis is true; (2) a random match (RM) occurred; or (3) a false positive match occurred (FP).

Suppose the random match probability for the DNA match evidence is rather low, say 10^{-9}, and there is no uncertainty associated with this number. Consider now two ways of assessing the DNA match. First, disregarding the possibility of a false positive—setting FP to 0—makes the match evidence appear extremely strong. In this case, the minimal prior sufficient for the posterior to be above .99 is only 0.001, where the relation between the prior probabilities and the posterior probabilities of the source hypothesis is given by the dashed orange line in Figure 11. What happens after taking into account the possibility of a false positive match? This depends on how this possibility of error is quantified. Assume the false positive rate corresponds to the upper bound of the [0.001, 0.01] interval. This assumption completely changes the assessment of the match evidence. Now the posterior of .99 is reached only if the prior is above .99. The match evidence appears to be extremely weak. So which is it? As already seen in the introduction, the point estimate exaggerates the value of the match evidence, while using the upper bound of the false positive rate has the opposite effect. What happens within the [0.001, 0.01] interval cannot be ignored.

To take into consideration the values within the edges, it would be best to have a good density estimate of the false positive errors frequency, as we should, if the issue had been properly studied. But we do not. For now, we will illustrate the consequences of taking two different approaches. On one approach, any value between the edges is considered equally likely (and we add a little leeway on top). On another approach, not all values are equally likely—for example, suppose you think it is 50% likely that the false positive rate is below .0033. In addition, suppose the distribution, while being centered closer to zero, is long-tailed (we used a truncated normal distribution here). These two distributions are displayed in Figure 9. On both approaches, we assume that the false positive rate is between 0.001 and 0.01 with 99% certainty. The uniform distribution—which regards all false positive rates in the interval as equally likely—leads to a rather conservative evaluation of the match evidence, much more so than the truncated normal distribution. This is apparent from Figure 10 and 11, which show the prior probabilities of the source hypothesis needed to secure a posterior probability above .99. Working with a distribution—more so if it is not a uniform distribution—affords a more balanced assessment of the evidence than simply relying on the edges of an interval.

The lingering question, however, is how these distributions can be obtained. Admittedly, studies on false positives are limited and only give an incomplete picture. More studies are needed. This does not mean, however, that until then using point estimates and interval edges is preferable. After deciding on the functional form of a distribution—such as truncated normal or beta—only a few numbers need to be elicited from experts for constructing a density.[23] Having to rely on such elicitation is not without problems, but it is better than asking experts for single point estimates and relying on these (O'Hagan et al., 2006).

[23] For instance, assuming the distribution is a truncated normal, it is enough for the expert to assert that both the 99% interval is as the one we used, and that they believe with more than 50% confidence the false positive rates to be below .033 for the curve to be determined.

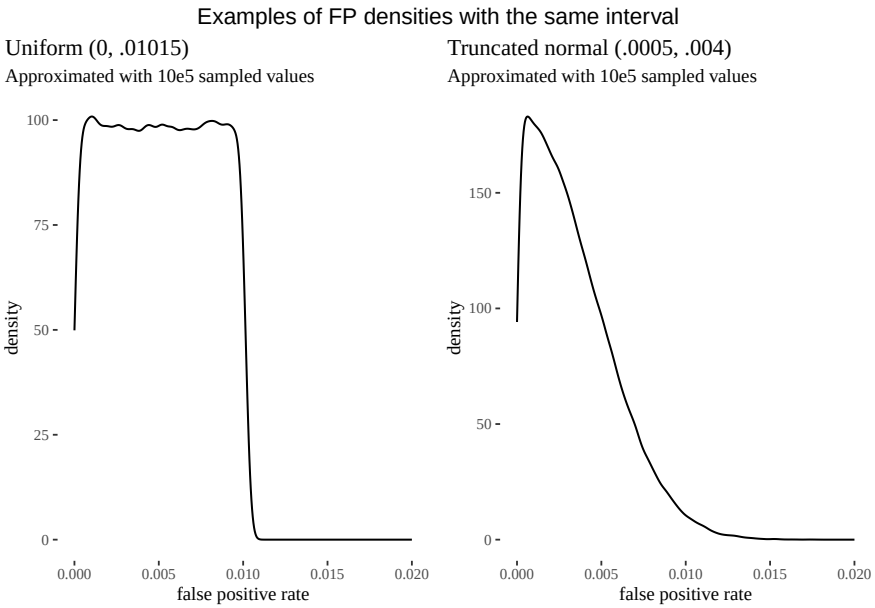

Figure 9: Two examples of assumptions about the false positive rates, both having pretty much the same 99% highest density intervals. Left: all error rates are equally likely. Right: the most likely values are closer to 0, but also some high values while unlikely are possible.

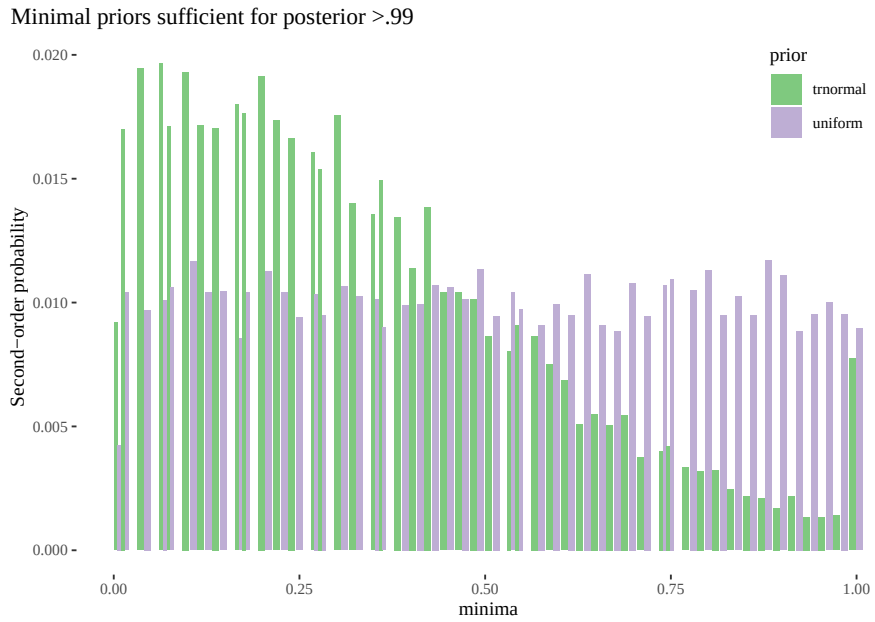

Figure 10: The distribution of minimal priors sufficient for obtaining a posterior above .99 on the two distributions of false positive rates. The truncated normal distribution has its bulk towards the left, but at the same time has higher ratio of evens in which this posterior is never reached.

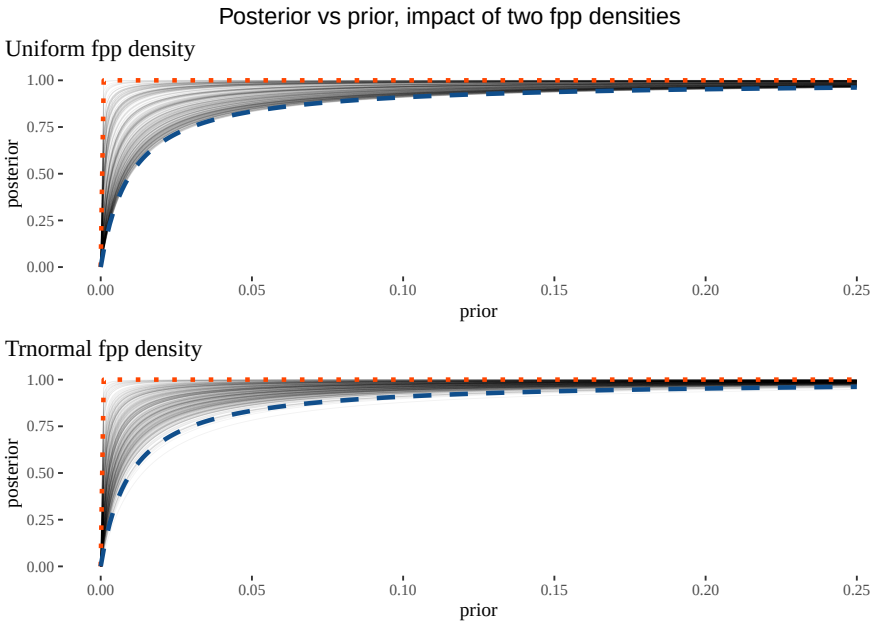

Figure 11: Impact of prior on the posterior assumign two different densitites for false positive rates. Note how both the "pristine" error-free point estimate (orange) and the charitable version (blue) are quite far from where the bulks of the distributions in fact are. Note also how the trnormal density allows for even more charitable cases, which results from it being long-tailed.

4.2 Higher-order Bayesian Networks

The higher-order framework we are advocating is not only applicable to the evaluation of individual pieces of evidence. Complex bodies of evidence—for example, those represented by Bayesian networks—can also be assessed using higher-order probabilities. One fairly straightforward way to go about this is to stochastically generate Bayesian networks using our uncertainty about the parameter values, update with the evidence, and propagate uncertainty to approximate the marginal posterior for nodes of interest.

As an illustration, let us start with a simplified Bayesian network developed by Fenton & Neil (2018). The network is reproduced in Figure 12 and represents the key items of evidence in the infamous British case R. v. Clark (EWCA Crim 54, 2000).[24]

In a Bayesian network the arrows depict direct relationships of influence between variables, and nodes—conditional on their parents—are taken to be independent of their non-descendants. Amurder and Bmurder are binary nodes corresponding to whether Sally Clark's sons, call them A and B, were murdered. These nodes influence whether signs of disease (Adisease and Bdisease) and bruising (Abruising and Bbruising) were present. Also, since A's death preceded in time B's death, whether A was murdered casts some light on the probability that B was also murdered.

The choice of the probabilities in the network is quite specific, and it is not clear where such precise values come from. The standard response invokes *sensitivity analysis*: a range of plausible values is tested. As already discussed, this approach ignores the shape of the underlying distributions. Sensitivity analysis does not make any difference between probability measures (or point estimates) in terms of their plausibility, but some will be more plausible than others. Moreover, if the sensitivity analysis is guided by extreme values, these might play an undeservedly strong role. These concerns can be addressed, at least in part, by recourse to higher-order probabilities. In a precise Bayesian network, each node is associated with a probability table determined by a finite list of numbers (precise probabilities).

[24] Sally Clark's first son died in 1996 soon after birth, and her second son died in similar circumstances a few years later in 1998. At trial, the pediatrician Roy Meadow testified that the probability that a child from such a family would die of Sudden Infant Death Syndrome (SIDS) was 1 in 8,543. Meadow calculated that therefore the probability of both children dying of SIDS was approximately 1 in 73 million. Sally Clark was convicted of murdering her infant sons. The conviction was reversed on appeal. The case of appeal was based on new evidence: signs of a potentially lethal disease were found in one of the bodies.

But suppose that, instead of precise numbers, we have densities over parameter values for the numbers in the probability tables.[25]
An example of a higher-order Bayesian network for the Sally Clark case is given in Figure 13.

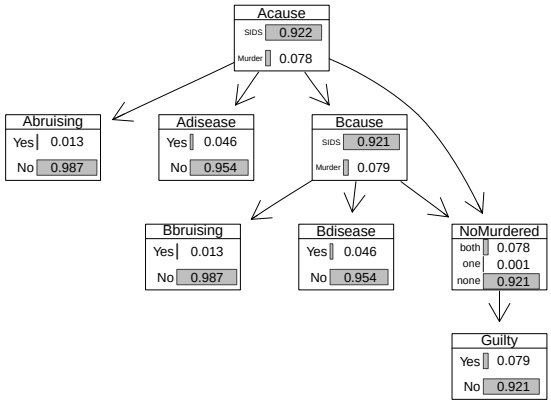

Figure 12: Bayesian network for the Sally Clark case, with marginal prior probabilities.

[25]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. In what follows we will work with sample sizes of 10k.

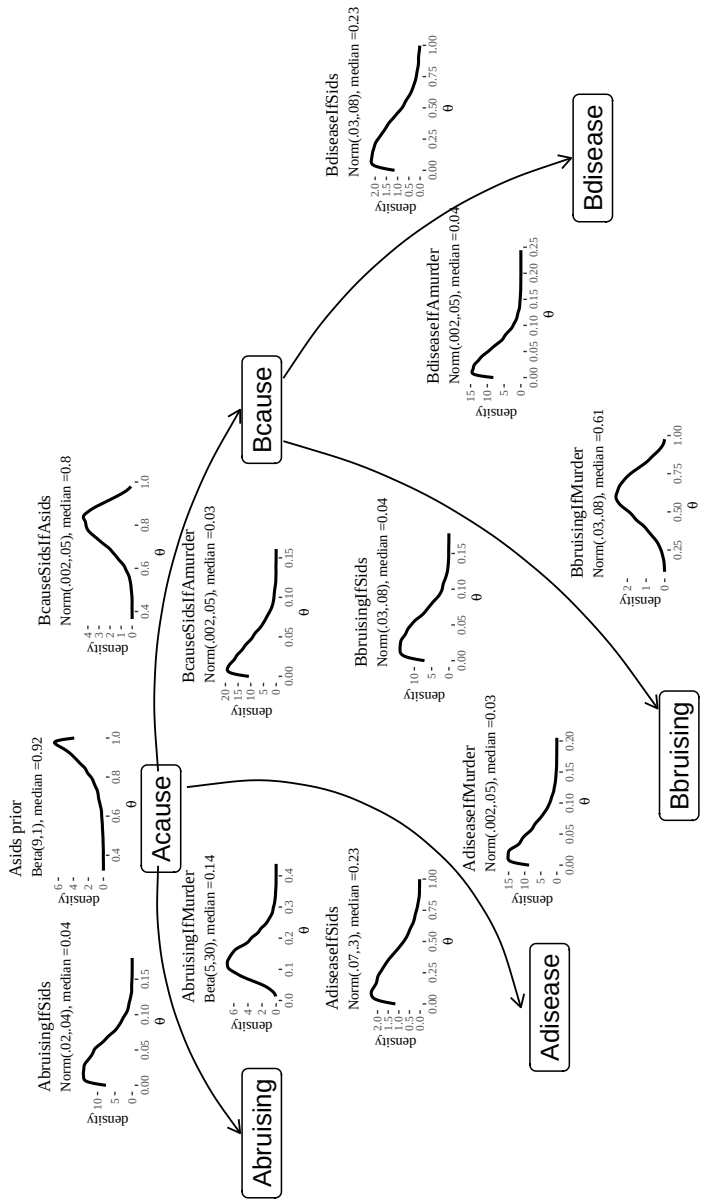

Figure 13: Example of a higher-order Bayesian network for the Sally Clark Case.

With the help of the higher-order Bayesian network, we can investigate the impact of different items of evidence on Sally Clark's probability of guilt (Figure 14). The starting point is the prior density 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 on the first child. Interestingly, in the strongest scenario against Sally Clark (third graph), the median of the posterior distribution is above .95, but the uncertainty around that median is still too wide to warrant a conviction.[26] This underscores the fact that relying on point estimates can lead to overconfidence. Paying attention to the higher-order uncertainty about the first-order probability can make a difference to trial decisions.

[26]The lower limit of the 89% Highest Posterior Density Intervals (HPDI) is at .83.

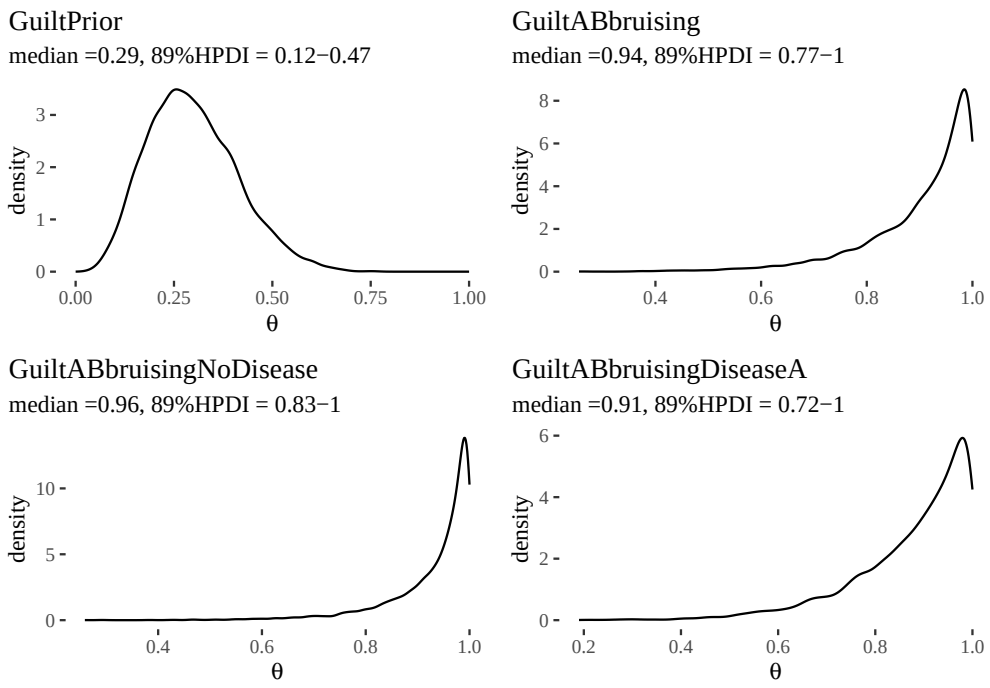

Figure 14: Impact of incoming evidence in the Sally Clark case.

One question that arises is how this approach relates to the standard method of using likelihood ratios to report the value of the evidence. On this approach, the conditional probabilities that are used in the likelihood ratio calculations are estimated and come in a package with an uncertainty about them. Accordingly, these uncertainties propagate: to estimate the likelihood ratio while keeping track of the uncertainty involved, we can sample probabilities from the selected distributions appropriate for the conditional probabilities needed for the calculations, then divide the corresponding samples, obtaining a sample of likelihood ratios, thus approximating the density capturing the recommended uncertainty about the likelihood ratio. Uncertainty about likelihood ratio is just propagated uncertainty about the involved conditional probabilities. For instance, we can use this tool to gauge our uncertainty about the likelihood ratios corresponding to the signs of bruising in son A and the presence of the symptoms of a potentially lethal disease in son A (Figure 15).

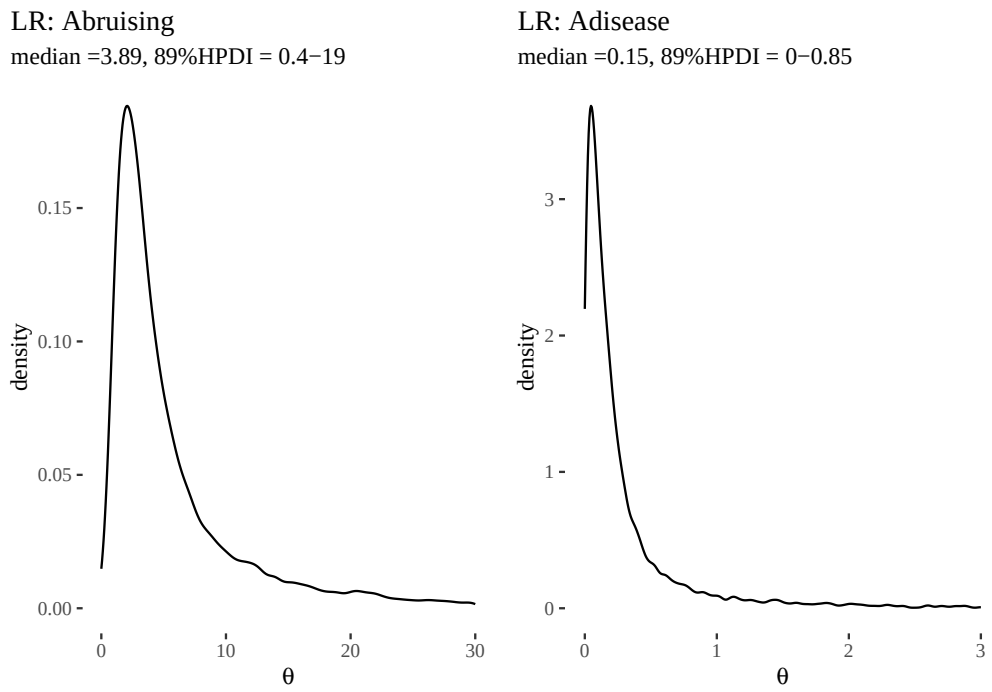

Figure 15: Likelihood ratios forbruising and signs of disease in child A in the Sally Clark case.

4.3 Relationship with Bayesian hierarchical models

Our approach does involve multiple parameters, uncertainty about them, and some dependency structure between random variables. So it is only natural to ask whether what we propose is not just an old wolf in a new sheep's clothing, as one might think that what looks like a DAG and quacks like a DAG is always a hierarchical model. As this question has been brought up to us a couple of times in discussion, we decided it deserves an elaboration.

First, we need some clarity on what a Bayesian hierarchical model is. In the widest sense of the word, these are mathematical descriptions involving multiple parameters such that credible values for some of them meaningfully depend on the values of other parameters, and that dependencies can be re-factored into a chain of dependencies. For instance, think about a joint parameter space for two parameters θ and ω, where $p(\theta, \omega|D) \propto p(D|\theta, \omega)p(\theta, \omega)$. If, further some independency-motivated re-factoring of the right-hand side, for instance as $p(D|\theta)p(\theta\omega))p(\omega)$ is possible, we are dealing with a hierarchical model in the wide sense of the word.

Such models come useful when we are dealing with clustered data, such as a cohort study with repeated measures, or some natural groupings at different levels of analysis. Then, lower-level parameters are treated as iid and share the same parameter distribution characterized by some hyper-parameters in turn characterized by a prior distribution. As a simple example consider a scenario in which we are dealing with multiple coins created by one mint—each coin has its own bias θ_i, but also that there is some commonality as to what these biases are in this mint, represented by a higher-level parameter θ.

keep k fixed – then ω is our expected value of the θ_i parameters, with some variety around it determined by k. Now, if we also are uncertain about ω and express our uncertainty about it in terms a density $p(\omega)$ we got ourselves a hierarchical model with joint prior distribution over parameters $\prod p(\theta_i|\omega)p(\omega)$. As an another example, one can develop a multilevel regression predicting, say the distributions of the radon levels in various counties, where both the intercept and the slope vary with counties by taking

fix the friggin formulas in the file

How is our approach different. Conceptually, at a few levels. For one, we are not dealing with clustered data. Given a hypothesis H and an item of evidence E for it, where both are propositions (or binary random variables), there seems to be no interesting conceptualizations on which the underlying data would be clustered. For conceptual reasons, it is not that stains at a crime scene are a subgroup of crimes being committed. Yes, there is dependency between these phenomena, but describing it as clustering would be at least misleading. Second, the dependencies proceed through the values of the

random variables and require also conditional uncertainties regarding the dependencies between the variables.

Fix formulae, see source

Yes, there is a meaningful independence/dependence structure here, but the re-factoring in terms of the actual values of the random variables makes it quite specific, at the same time allowing for the computational use of Bayesian networks. Finally, the reasoning we describe is not regression the way it is normally performed: the learning task is delegated to the bottom level of whatever happens to the Bayesian networks once updated with evidence.

5 Weight of Evidence

M's comments in boldface.

M's comment: The discussion about weight seems the start of a new paper. We should probably work on the previous sections and make them more defensible and better grounded in the literature.

After sketching how the legal applications of the higher-order approach should go, we turn to another payoff of higher-order probabilism: it allows us to develop a theory of the weight of evidence that outperforms the existing proposals. We will start with an informal sketch of the concept of the weight of evidence, as opposed to the balance of the evidence. We will then explore a few attempts at modeling this idea, first from the precise's and then from the imprecise's perspective. Finally, we will show that higher-order probabilism can offer a better theory.

5.1 Examples and Desiderata

In the 1872 manuscript *The Fixation of Belief* (W3 295), C. S. Peirce makes the following observation about sampling from a bag of beans that are either black or white:

> When we have drawn a thousand times, if about half have been white, we have great confidence in this result ... a confidence which would be entirely wanting if, instead of sampling the bag by 1000 drawings, we had done so by only two.

In both cases, our best assessment of the probability that the next draw will be a black bean is .5, but how sure we should be of that assessment is quite different depending on whether it is based on two or one thousands draws. In other words, the *weight* of the evidence seems much greater after drawing a thousand beans and finding out that half are black, compared to drawing just two beans and again finding out that half are black. The weight of the evidence is different in the two cases, but its *balance*—understood here as the empirical proportion of black-to-white beans—is the same.[27]

Peirce did not use the expression the weight of evidence (and, in fact, he used the phrase to refer to the balance of evidence, W3 294) (Kasser, 2016). However, his remarks anticipated what came to be called weight of evidence by Keynes in his 1921 *A Treatise on Probability*:

> As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either increase or decrease, according as the new knowledge strengthens the unfavourable or the favourable evidence; but something seems to have increased in either case—we have a more substantial basis upon which to rest our conclusion. I express this by saying that an accession of new evidence increases the weight of an argument. (p. 71)

The key point is the same (Levi, 2011): since the balance of probability alone cannot characterize all important aspects of evidential appraisal, another dimension—the weight of an argument—must be deployed to quantify uncertainty.[28]

It is instructive to examine more closely Keynes' claim that the weight of evidence, unlike its balance, always increases as more evidence is taken into account. While balance can oscillate one direction or the other, weight would seem to always increase. We can state this requirement as follows:

[27] Similar remarks can be found in Peirce's 1878 *Probability of Induction*. There, he also proposes to represent uncertainty by at least two numbers, the first depending on the inferred probability, and the second measuring the amount of knowledge obtained. For the latter, Peirce proposes to use some dispersion-related measure of error (but then suggests that an error of that estimate should also be estimated and so on, so that ideally more numbers representing errors would be needed).

[28] Keynes entertained the possibility of measuring weight of evidence in terms of the variance of the posterior distribution of a certain parameter, but was quite attached to the idea that weight should increase with new information, even if the dispersion may increase with new evidence [TP 80-82]. So he proposed only a very rough sketch of a positive proposal. Moreover, he was unclear how a measure of weight could be part of decision-making. He was ultimately skeptical about the practical significance of the notion [TP 83].

(Monotonicity) If E is relevant to X given K, where K is background knowledge, $V(X|K \wedge E) > V(X|K)$, where V is the weight of evidence.

Monotonicity is consistent with Peirce's example involving drawing from a bag of beans. As the sample size increases and the relative proportion of black-to-white beans remains constant, the weight of the evidence increases. That is:

(Weak Increase) In urn-like cases, the evidential weight obtained by a larger sample is greater, if the relative frequencies in the samples remain the same.

This formulation can be strengthened by dropping the assumption of equal relative frequencies:

(Strong Increase) In urn-like cases, the evidential weight obtained by a larger sample is higher.

We think there are good reasons to reject Monotonicity and Strong Increase, but we agree with Weak Increase. Monotonicity and Strong Increase are consistent with a certain conception of the weight of evidence, what we might call *quantity* of evidence. There is no doubt that, as more evidence is taken into account, the quantity of the evidence must increase. But the weight of evidence need not be identified with its quantity alone. As more evidence is taken into account, the new evidence may speak less clearly in favor or against a hypothesis. For consider this example:

> **A (possibly) rigged lottery:** Initially, you think the lottery is fair. You have no reason to doubt that. So, you calculate precisely the probability that a certain ticket number will be drawn. Then, rumors begin to surface that the lottery is rigged and that only numbers that satisfy a complicated equation will be drawn. You now have more relevant evidence at your disposal, but that evidence is more confusing and muddled than before.

Arguably, this is a scenario in which the quantity of evidence has increased, but the weight of evidence has not. If this is right, weight and quantity can come apart.

So, we seek a theory of evidential weight that can model two intuitions. The first is that, as the sample size increases, the weight of the evidence must also increase (under certain conditions, along the lines of Weak Increase). The second intuition is that that, even when the quantity of evidence increases, the weight of evidence might not (as illustrated by the example of the rigged lottery).

5.2 Weight and Precise Probabilism

An obvious place to look for a theory of the weight of evidence is within precise probabilism. The earliest account of the weight of evidence is given by Good (1950), as follows:

$$W(H : E) = \log \frac{\mathsf{P}(E|H)}{\mathsf{P}(E|\neg H)},$$

where E is the evidence and H a hypothesis of interest. This account follows naturally from the assumption that that $W(H : E)$ must be a function of $\mathsf{P}(E|H)$ and of $\mathsf{P}(E|\neg H)$. As Good (1985) put it, "I cannot see how anything can be relevant to the weight of evidence other than the probability of the evidence given guilt and the probability given innocence" [p 250].[29]

One important question is whether Good's weight satisfies the desiderata we already discussed. We can investigate further developing Good's own example. If, in an experiment, the observations $E_1, \ldots, E_K$ are independent given H and given $\neg H$, the resulting joint likelihood is the result of multiplying the individual likelihoods. Thus, the joint weight is the result of adding the individual weights. Now, suppose a die is selected at random from a hat containing nine fair dice and one loaded die with bias $1/3$ of obtaining a six. Every time you throw the die and obtain a six, the weight for the hypothesis that the die is biased increases by $log_{10}(\frac{1/3}{1/6}) = log_{10}(2)$, that is 0.30103, and every time you throw it and obtain something else, the weight changes by $log_{10}(\frac{2/3}{5/6}) = log_{10}(.8)$, that is -0.09691. The weights in db (that is, multiplied by 10) for all possible outcomes of up to 20 tosses are displayed in Figure **??**.

Hey, why wasn't it here in the first place? Can you reinstate it, Marcello?

[29] We can: how good those estimates are and what our uncertainty about them is. And if you want to have a context-relative notion of weight of evidence, so that weight considerations tell you when further investigation is undesirable as the potential weight of evidence is not that high given what you already know, also the weight of evidence and the posterior resulting from the evidence you have obtained so far.

Two facts are notable. First, Good's weight can drop with sample size. For instance, the weight for 5 sixes and 4 other numbers is 1.2db, and it is .2db for 5 sixes and 5 other numbers. In addition, weight can drop while the sample size increases even if the proportion of sixes remains the same. For instance, if none of the observations are sixes, the weights go from -10 to -19.7 as the sample size goes from 0 to 10. Less trivially, the observation of one six in five leads to weight of -10.9, while the observation of two sixes in ten tosses leads to weight -11.7. That is, (Monotonicity) (Weak Increase) and (Strong Increase) all fail on Good's measure. This suggests measure is too closely connected to likelihood ratio for our purpose and as such does not capture the notion of weight of evidence that we are after.

One way to think about this feature of Good's weight is that the (log of the) likelihood ratio is a *directional* measure: the evidence may be favorable or unfavorable to a hypothesis (compared to another). By contrast, Keynes' remarks suggests that weight is a non-directional measure. It appears to always increase no matter the balance (though we will ultimately reject this claim trying to find a middle ground between these approaches).

So one might think that all there is to be done is stripping the likelihood ratio of its directionality, say by taking the absolute value of the natural log (Nance, 2016, sec. 3.5). The weight of evidence E relative to the pair of hypotheses H, H' would be

$$|\ln(LR_{H,H'}(E))|,$$

where $LR_{H,H'}(E) = \frac{P(E|H)}{P(E|H')}$. By the properties of logarithms, $ln(1/x) = -ln(x)$, so two items of evidence of equivalent strength—but opposite directionality—would have the same weight. So, for example, $|ln(1/3)| = |ln(3)| = 1.61$.

This account, while simple and elegant, faces difficulties. As it is ordinally equivalent to the standard ordering of the absolute values of Good's weights, our objections, mutatis mutandis, apply. Moreover, as Nance points out, there is a decomposition problem. Consider two items of evidence that, taken together, have a likelihood ratio of one, say one has a likelihood ratio of 1/3 and the other of 3. Assuming they are probabilistically independent given a hypothesis of interest, their combined likelihood ratio results from multiplying the individual likelihood ratios. Thus, their combined weight would be zero since $\ln(LR_H(E1 \wedge E_2)) = \ln(1) = 0$. However, by adding the weights one by one, the combined weight would be different from zero, since $|ln(1/3)| + |ln(3)| = 3.22$. So, do the two items of evidence have zero weight or not? Depending on how evidence is decomposed, it appears to have different weights.

To think about: does the decomposition problem apply to Good's weight?

5.3 Weight and Imprecise Probabilism

Precise probabilism was not very successful at delivering an account of weight. Does imprecise probabilism fare any better? Weatherson (2002) proposes the following: a body of evidence is weightier whenever the representor set of probability measures compatible with the evidence is, in some sensible sense, smaller. Consider a case of complete ignorance about the bias of a coin. The representor set will contain *any* probability measure. This corresponds to complete lack of evidence and null weight, as expected. At the other extreme, consider a case in which the fairness of the coin is known for sure. The supporting evidence here would have maximal weight and the representor set would only contain the precise probability measure that assigns .5 to the two outcomes. All other intermediate cases would fall somewhere in between.[30]

M: the decomposition problem does not apply to Good who has a directional notion of weight.

R: the bit I commended out was false, I think. See the revised verson of the above passage, and think hard about the argument

This account accommodates the intuition underlying the rigged lottery example. As more evidenced is accumulated, the representor set can become larger and include more probability measures than before. Thus, weight of evidence can decrease even when the quantity of evidence increases. This is promising. The problem, however, is that this theory of weight inherits the difficulties of imprecise probabilism, which we have already mentioned. For example, recall that imprecise probabilism cannot model a situation in which an epistemic agent believes that the coin could have bias .8 or .4, but thinks that one bias is more likely than the other. Arguably, evidence compatible with the coin having two equally likely biases should possess different weight than evidence compatible with the coin having two biases one more likely than the other. The latter situation is closer to full weight—the situation in which the exact bias of the coin is known for sure. But a theory of weight based on imprecise probabilism is ill-equipped to accommodate these nuances.

M: Why was it false?

In what sense? How is this defined?

R: I reinstated but abridged a discussion of Peden and Joyce

[30] Let x_1 and x_2 the two extreme probability assignments in the representator set compatible with the available evidence E. Then, the weight of E would be 1-x_1 − x_2. As expected, if there is only one probability measure, the weight would be one.

There are two proposals on the market that are somehow related to imprecise probabilism. One, due to Peden (2018) follows a suggestion from (Kyburg, 1961). Kyburg proposed using the degree of imprecision of the intervals in his probability system called Evidential Probability (EP), and a range of rules describing how such intervals are to be shaped by the evidence. With this system in the background, Peden proposes the following definition of the weight of the argument for H given E and K, where $\mathsf{EP}(H|E \wedge K) = [x, y]$:

$$\mathsf{WK}(\mathsf{H}|\mathsf{E} \wedge \mathsf{K}) = 1 - (y - x) \tag{WK}$$

That is, the weight of the evidence is the spread of the evidential probability, transformed to scale between 0 and 1, reaching 1 when the spread is 0 and 0 when the spread is 1. A similar line is taken within standard imprecise probabilism by Walley (1991), who proposes to take the edges of the resulting interval that captures changes in the weight of evidence.

We already criticized the focus on the edges of the intervals in this paper. Relatedly, on this proposal the edges of the intervals are what contributes to WK, and these are highly sensitive to the choice of the margin of error, but what margin of error to choose and why remains a mystery, and what margin of error has been chosen does not function anywhere in the EP representation of uncertainty. It may easily happen that for two different distributions the 1% intervals will be identical while the 78% intervals will not. Such differences will obviously not be captured by the 1% margin of error intervals.

Another stab at explicating weight of evidence within the IP framework has been made by Joyce (2005). Joyce uses a density over chance hypotheses to account for the notion of evidential weight. He conceptualizes the weight of evidence as an increase of concentration of smaller subsets of chance hypotheses:[31]

$$w(X, E) = \sum_x |c(ch(X) = x|E) \times (x - c(X|E))^2 - c(ch(X) = x) \times (x - c(X))^2| \tag{Joyce}$$

The idea here is that weighty evidence should make the credence resilient, and resilience makes the difference between the posterior credence in chances $c(ch(X) = x|E)$ and the prior credence in chances $c(ch(X) = x)$. The complication is that the impact of this difference should be lower for for those values of x that are close to $c(X|E)$ for the posterior and close to $c(X)$ for the prior. Hence, the formula for w is takes (the absolute value of) the difference between posteriors and priors weighed by, these (squared) distances. The weightier the evidence, the smaller w is supposed to be.[32]

There are various issues with this approach. One is that now to evaluate the weight of evidence E with respect to proposition X now you need to have and use in your calculations your estimation of chances of X. Let us put aside the worry that it is not obvious that we can meaningfully talk about chances of arbitrary propositions. Even then, the name of the game for the imprecise probabilist was to express the uncertainty about X in terms of a representor, a set of probability measures. However, one can have a representor with respect to a set of object-level propositions including X without having a

[31] This looks a bit complicated, so let us take a slow look at an example. Suppose you only consider three chance hypotheses, that the coin bias is one of .4, .5, and .6, that is, the hypotheses are $ch(X) = .4, ch(X) = .5$, and $ch(X) = .6$. For each $x \in \{.4, .5, .6\}$) you attach a prior credence $c(ch(X) = x)$ to the corresponding hypothesis. Say you start with equal priors, that is for all $x \in \{.4, .5, .6\}$ you have $c(ch(X) = x) = 1/3$. Then, your expected value of X, which Joyce takes to be your credence in X simpliciter is $\sum_x c(Ch(X) = x)x$, which is .5.

Now consider your evidence: you tossed the coin and observed, say, seven heads out of ten tosses. We need $c(ch(X) = x|E)$. By Bayes, we have:

$$c(ch(X) = x|E) = \frac{c(E|ch(X) = x)c(ch(X) = x)}{c(E)},$$

so we need to calculate the likelihoods, $c(E|ch(X) = x)$. We assume you are probabilistically coherent, that you defer to chances, and know the experimental setup, so that the likelihoods are calculated using the binomial distribution, i.e. if the evidence is a heads and b tails:

$$c(E|ch(X) = x) = \binom{a+b}{a} x^a (1 - x)^b$$

In our example, the likelihoods (rounded) are .042, .117, and .214 respectively. The denominator is calculated by taking $c(E) = \sum_x c(E|ch(X) = x)c(ch(X) = x)$, which in our case turns out to be .124. Putting these together, the values of $c(ch(X) = x|E)$ are .113 .312, and .573 (rounded). Then, your expected value, which Joyce to be your credence in X simpliciter conditional on E is $\sum_x c(Ch(X) = x|E)x$, which is .54.

[32] Accordingly, in our example the weights for the prior are $-.1^2, 0, .1^2 = 0.01, 0, .01$, the weights for the posterior are .021330539, .002120582, and .0029 and w is 0.003241822. For comparison, if instead we observed 70 heads in 100 tosses, w would be .006689603.

single credence about chances, so now the calculations of weight of E with respect to X do not fall out whatever was supposed to capture the agent's uncertainty about X, E and their relationship.

Crucially, the measure displays strange behaviors when run on even straightforward cases. Expanding a bit on an example that we've been discussing in a footnote, observe that measure might result in drastic shift in weights even if the observed frequencies are not too far from the chance hypotheses, and that the weight of evidence might drop as the sample size increases, even if the observed success frequency remains the same, which means that (Weak Increase) fails for this measure (see Figure 16).[33]

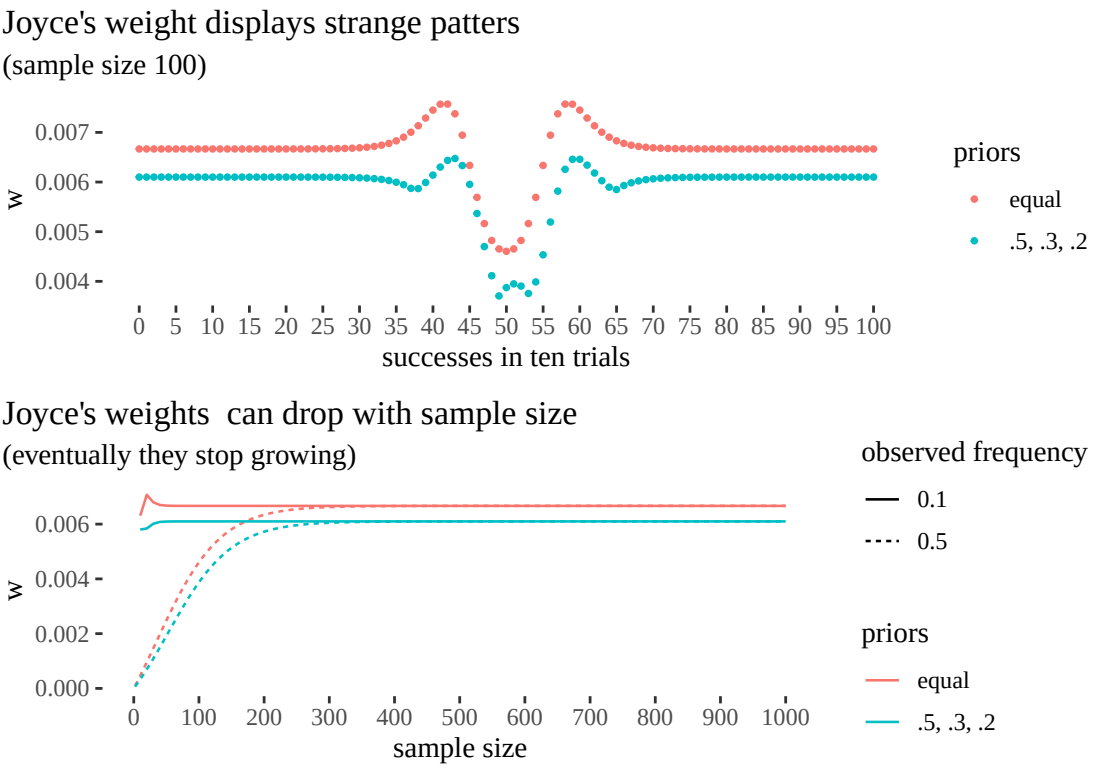

Figure 16: (top) Joyce's w (the lower it is, the higher the weight) for various observed successes in 100 Bernoulli trials. Three chance chypotheses: $.4, .5, .6$, and two sets of priors: equal and $.5, .3, .2$ respectively. Agan, the weightiest evidence is obtained with successes close to the expected value, with large variation for observed frequencies not too far from the expected values, fairly flat otherwise. (bottom) Joyce's w (the lower it is, the higher the weight) for two fixed success ratio across various observed successes in Bernoulli trials (lines are used for smoothing). Note large shifts with possible decrease in the beginning, and a flattening afterwards.

Moreover, the approach is of limited applicability. For one thing, as Joyce admits, it is supposed to work when RA's credence is mediated by chance hypotheses. Depending on applications, such a mediation might be unavailable. Another issue is that this might work for unimodal distributions

[33] What is the reason for this strange behavior? The shaping of Joyce's weight is a balancing act. For instance, for frequency .1 with equal priors the weight is maximized at $n = 90$ and starts dropping at $n = 100$. Why? We start with three chance hypotheses, $.4, .5, .6$ with equal priors. Once the observations have been made, the posterior for the chance hypotheses is focused at the chance hypothesis .4 (its posterior is more or less .99999), and so is the credence in X simpliciter (this expected value is .4000003). Now the weights are build by measuring square distance from the credence in X simpliciter and since the expected value is nearly equal to the lowest chance hypothesis under consideration, the weight for the lowest chance hypothesis is $8.260660e - 14$, so while the posterior for this hypothesis is very high, the weight is very low and its contribution to the weight calculation is severely limited. Ultimately, what happens with weight is now a matter of balancing the uneven posteriors with squared penalties (or rewards, really) for the distance from the expected value (which is pretty much the most likely hypothesis once you have made enough observations). Once you observe 10 successes in 1000 trials, the credence in X simpliciter becomes .4000001, so the distance of the lowest chance hypothesis to it drops, and this weight drops "faster" than the resulting increase in the probability of the lowest chance hypothesis itself. The stabilization is achieved because further on the posterior for this hypothesis can only get closer to one (and closer to zero for all the other hypotheses).

when we only consider the influx of new data points, but it's unlikely to give desired results if, say, the evidence obtained is the testimony of disagreeing witnesses. This is because an essential part of the calculations relies on taking the expected value, and it is not too hard to imagine cases of diverging items of evidence resulting only in a small chance of the expected value. This is related to the assumption that an agent's stance towards a proposition should be represented as the expected value of the chance hypothesis—which we already extensively argued against.

Finally, note that the proposal employs probabilities or probability densities (if we go continuous) over parameter values. This is in line with what we propose if we do not assume these are chances and treat them as, say, parameters that are potentially rational to accept in light of the evidence. But then, there are useful ways to go this way without turning to IP, [34] to which we now proceed.

5.4 Weight and Higher-order Probabilism

We now present our own account of the weight of evidence. This account has two distinctive features. First, it is based on higher-order probabilism. Second, it is information-theoretic. To develop this account, we will begin with a short introduction to Shannon's theory of information.

5.4.1 Entropy of a Distribution

Let X be a random variable and P a probability distribution over its values.[35] Shannon's measure of information, $H(X)$, reads:

$$H(X) = -\sum \mathsf{P}(x_i) \log_2 \mathsf{P}(x_i)$$

Consider the simple case in which X can take two values—outcome 1 and outcome 0—whose probabilities are p and $1-p$. Figure **??** shows $H(X)$ as a function of p. Entropy $H(X)$ is greatest when the two outcomes have an equal probability of .5. The more the probabilities deviate from .5, the more $H(X)$ approaches zero. To make sense of this, $H(X)$ can be thought as the *entropy* (i.e. the lack of information) contained in the distribution associated with random variable X. When the two outcomes have equal probability, entropy is greatest. When they have different probabilities, one outcome will be more probable than the other. The more probable one of the outcomes (and thus the less likely the other), the lower the entropy. Intuitively, $H(X)$ captures the idea that entropy is greatest when the indecision about which outcome will occur is maximal, and the entropy decreases when such indecision decreases.

Add plot of H(X)

If $H(X)$ is a measure of entropy—that is, a measure of lack of information—why call it a measure of information? $H(X)$ is also a measure of information in the following sense: it describes the *expected amount of information* one would receive upon learning the actual value of X. After all, the higher the entropy, the less informative a distribution, the more you expect to learn upon finding out the actual value of X. Conversely, the lower the entropy of the distribution, the more informative the distribution, the less one expects to learn upon finding out the actual value of X.

5.4.2 Weight of a Distribution

Since $H(X)$ can be thought as the entropy of a distribution, we will switch to the notation $H(P)$. This notation emphasizes the distribution P rather than the random variable X. The entropy of a distribution is to be contrasted with its informativeness, which we will denote by $W(P)$, the weight of the distribution. How should we measure the weight (or informativeness) of a distribution?

This temptation didn't make sense, entropy can often be higher than 1.

BEfore we proced, one technical remark. The move to continuous distributions is not straightforward,[36], so in what follows we prefer to stick to entropy proper and discretize. One reason is that we will want to meaningfully compare information conveyed by discrete distributions to that conveyed by continuous ones. A convenient way to do so is to abandon the idea that we should be infinitely precise, fix a certain number of bins (that is a certain level of precision) and keep it fixed in our comparison.

[34] After all, notice how the notion of a representor plays no role in Joyce's explication of weight whatsoever!

[35] Since working with continuous distributions is not straightforward, we will be using *grid approximations* of continuous distributions: we will split X into a 1000 bins and use the normalized densities for their centers to obtain their corresponding probabilities. As long as we do not change our level of precision (which would inevitably lead to changes in entropy) in our comparisons, this is not a problem.

[36] One might expect that entropy in the continuous case could be made by binning and taking the limit. For instance, suppose

This is what we will do: effectively, we will be using *grid approximations* of continuous distributions: we will split X into a 1000 bins and use the normalized densities for their centers to obtain their corresponding probabilities. As long as we do not change our level of precision (which would inevitably lead to changes in entropy) in our comparisons, this is not a problem. An additional advantage is that now we do not have to deal with the intricacies of explicit analytic calculations for continuous variables and comparing apples (entropy) with oranges (differential entropy). One side-effect of this is that the exact absolute values of entropy will depend on the levels of precision we choose, and so we should not be too attached to differences, focusin on proportions instead.

From this perspective, the weight (or informativeness) of a distribution is modeled by comparing it to the least informative distribution, the uniform distribution, which expresses complete uncertainty. The more informative a distribution, the more it departs from the uniform distribution, the more weight it has, on a scale from 0 to 1. If the drop from uncertainty is complete, the entropy drops to zero, and thus the weight should be 1; if the drop is null, the entropy remains the same, and thus the weight should be zero; if the drop is half, the weight should be .5; and so on for other intermediate cases. This pattern can be captured by the following definition of weight (or informativeness) of a distribution:

$$\mathsf{w}(\mathsf{P}) = 1 - \left(\frac{H(\mathsf{P})}{H(\mathsf{uniform})}\right)$$

where P is the probability distribution of interest and $\mathsf{uniform}$ is the baseline uniform distribution.[37]

This measure captures the two key intuitions we identified earlier. First, the weight of a distribution increases as the number of observations increases provided the relative frequency is fixed (Weak In-

---

we divide X into bins x_i of length Δ, so that we discretize X into X^Δ. The discrete case definition applies:

$$H(X^\Delta) = \sum \left[\mathsf{P}(X \text{ is in the } i\text{-th bin}) \log_2 \frac{1}{\mathsf{P}(X \text{ is in the } i\text{-th bin})}\right]$$

If you think of the histogram of the distribution of X^Δ with total area A, each bin has area a_i and height p_i. Suppose we normalize so that $A = 1$, then the probability of each bin is $\mathsf{P}_i = p_i\Delta$ and p_i can be thought of probability density. Then we have:

$$\begin{aligned}
H(X^\Delta) &= \sum \mathsf{P}_i \log_2 \frac{1}{\mathsf{P}_i} \\
&= \sum p_i\Delta \log_2 \frac{1}{p_i\Delta} \\
&= \sum \left[p_i\Delta \left(\log_2 \frac{1}{p_i} + \log_2 \frac{1}{\Delta}\right)\right] \\
&= \sum p_i\Delta \log_2 \frac{1}{p_i} + \underbrace{\sum \underbrace{p_i\Delta}_{\mathsf{P}_i}}_{1} \log_2 \frac{1}{\Delta} \\
&= \sum p_i\Delta \log_2 \frac{1}{p_i} + \log_2 \frac{1}{\Delta}
\end{aligned}$$

Accordingly, when we try to go continuous by taking the limit, we get:

$$H(X) = \left[\int_{-\infty}^{\infty} p(x) \log_2 \frac{1}{p(x)}\, dx\right] + \infty$$

This is as it should: the entropy of a continuous variable increases with the precision of measurement, so infinite precision gives infinite information. For this reason, for the continuous case it is usual to drop the rightmost part of the equation and talk about *differential entropy*:

$$\mathsf{H}(X) = \left[\int_{-\infty}^{\infty} p(x) \log_2 \frac{1}{p(x)}\, dx\right]$$

[37] Note that the entropy of a uniform distribution is pretty straightforward, so we can simplify:

$$\begin{aligned}
H(\mathsf{uniform}) &= \sum_{i=1}^{n} {}^1\!/_n \log_2 \frac{1}{{}^1\!/_n} \\
&= \log_2(n) \\
\mathsf{w}(\mathsf{P}) &= 1 - \left(\frac{H(\mathsf{P})}{\log_2(n)}\right)
\end{aligned}$$

crease; see Figure 17). As usual, we can think of a coin that is tossed a number of times and that lands heads or tails with a certain observed frequency. At the same time—and this is the second intuition—weight may diminish when the quantity of evidence increases, as the Rigged Lottery example suggest. Figure 17 shows that weight can drop—despite a larger number of observations—so long as the observed relative frequency changes from more extreme (say .1) to less extreme (say .5). That weight is not strictly tied to the quantity of evidence (number of observations) is also apparent from Figure 18. This shows that weight can vary dramatically depending on the observed relative frequency, all else being equal. The behavior of the proposed measure of weight can be illustrated more generally using distributions of various of shapes, displayed in Figure 19.

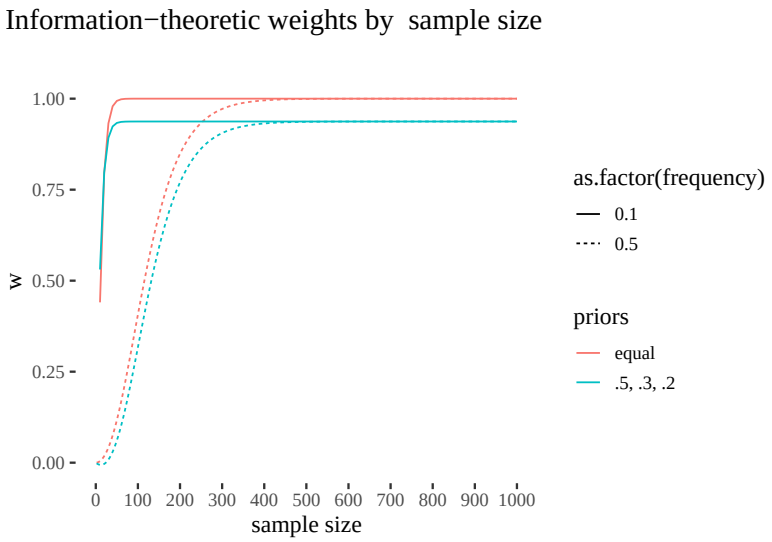

Figure 17: Entropy-based weight for two observed frequencies for various sample sizes (lines used instead of points for smoothing). Three chance chypotheses: $.4, .5, .6$, and two sets of priors: equal and $.5, .3, .2$ respectively.

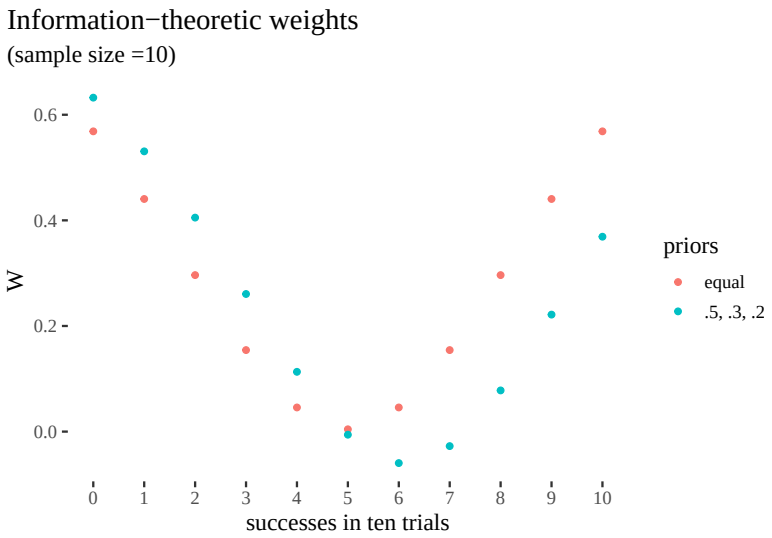

Figure 18: Entropy-based weight for for various observed successes in 10 Bernoulli trials. Three chance chypotheses: $.4, .5, .6$, and two sets of priors: equal and $.5, .3, .2$ respectively.

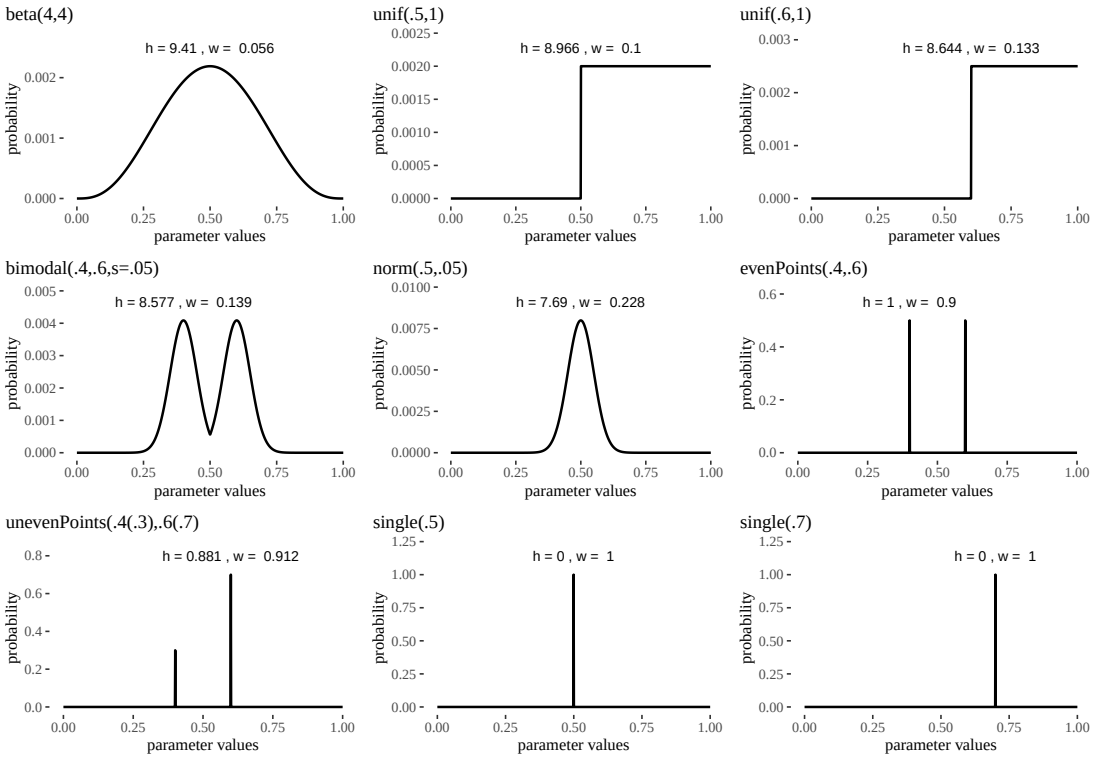

Figure 19: Examples of various distributions with their entropies and weights, ordered by weights. (1) beta(4,4), (2) uniform starting from .5 to 1, (3), uniform strating from .6 to 1, (4) two normal distributions centered around .4 and .6 with standard deviation .05, glued at .5. (5) normal centered around .5 with the same standard deviation, (6) one that assigns .5 to each of .4 and .6, (7) One that assigns .3 to .4 and .7 to .6., (8) one that assigns all weight to .5, and (9) one that assigns all weight to .7.

5.4.3 Weight of Evidence

So far we talked about the weight of a distribution, and indirectly the weight of the evidence so long as a distribution reflects the evidence. The notion of weight of evidence, however, be made more precise.

Suppose a distribution P depicts what an agent thinks, at some point in time, about the probability of the possible values of a random variable X, say the possible biases of a coin or first-order probabilities. In this sense, the weight $W(P)$ (or informativeness) of a distribution measures how informed an agent is about X. But it measures the information level of an agent only relative to the state of full uncertainty represented by the uniform distribution. There are two things missing from this account: one, not every agent starts with a uniform prior; two, how informed an agent is must depend on the evidence available to that agent. What we need, then, is an account of how informed agents are *on the basis of the evidence* they have, or in other words, an account of the weight (or informativeness) of the evidence they have.

If the agent starts with a uniform prior over the values of a random variable of interest, $W(P)$ would be a good enough approximation of how informed the evidence made them. In general, however, how much more information is obtained is context-dependent. The weight of evidence, then, must depend on what the agent already knows. Here is a general recipe. In a given context, consider the prior distribution P_0 for the target random variables X given what the agent already knows. Then, the agent updates by a body of evidence E. Call this posterior distribution P_E, where the updating is done by standard Bayesian conditionalization. Take the difference between the weight of the prior distribution, $W(P_0)$, and the weight of the posterior distribution, $W(P_E)$. The difference between the two–$W(P_E) - W(P_0)$ or more succinctly ΔW—measures the impact that evidence E has on the information level of the evidence.[38] The difference ΔW, then, is our proposed measure of the weight

[38] More precisely, the calculation follows the following schema:

of the evidence.[39]

Notice that our account of weight of evidence, in principle, is independent of the proposal to adopt second-order probabilities. The notions of $W(P)$ and ΔW could be applied to first-order probability distributions. Even if you just considered two competing hypotheses and the likelihood ratio, you could deploy our account of weight. But this would bring us back to Good's notion of weight, which does not capture the intuitions about weight of evidence that we wanted to capture. So it is crucial for our account of weight that it comprises two components: the information-theoretic and the higher-order component.

I will add a section on expected weight of evidence, and I didn't get the importance of the final challenge section, so I haven't copied it yet.

References

Bradley, D. (2015). *A critical introduction to formal epistemology*. Bloomsbury Publishing.

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*.

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

Dahlman, C., & Nordgaard, A. (2022). *Information economics in the criminal standard of proof*.

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.

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In A. Hajek & C. Hitchcock (Eds.), *Oxford handbook of philosophy and probability*. Oxford: Oxford University Press.

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

Elkin, L., & Wheeler, G. (2018). Resolving peer disagreements through imprecise probabilities. *Noûs*, *52*(2), 260–278. https://doi.org/10.1111/nous.12143

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

Good, I. J. (1950). *Probability and the weighing of evidence*. C. Griffin London.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics*. Elsevier Science.

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

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

Kasser, J. (2016). Two conceptions of weight of evidence in peirce's illustrations of the logic of science. *Erkenntnis*, *81*(3), 629–648.

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

Konek, J. (2013). *New foundations for imprecise bayesianism* (PhD thesis). University of Michigan.

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.

[39] If you prefer to think that weight of evidence should always be positive as a result of adding evidence, you might prefer the absolute value of the difference. We, however, prefer to keep track of whether the evidence makes the agent more or less informed about an issue.

Levi, I. (2011). The weight of argument. In *Fundamental uncertainty* (pp. 39–58). Springer.
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
Nance, D. A. (2016). *The burdens of proof: Discriminatory power, weight of evidence, and tenacity of belief*. Cambridge University Press.
O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., … Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*.
Peden, W. (2018). Imprecise probability and the measurement of keynes's "weight of arguments." *Journal of Applied Logics—IFCoLog Journal of Logics and Their Applications*, *5*(3).
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
Sjerps, M. J., Alberink, I., Bolck, A., Stoel, R. D., Vergeer, P., & Zanten, J. H. van. (2015). Uncertainty and LR: to integrate or not to integrate, that's the question. *Law, Probability and Risk*, *15*(1), 23–29. https://doi.org/10.1093/lpr/mgv005
Stewart, R. T., & Quintana, I. O. (2018). Learning and pooling, pooling and learning. *Erkenntnis*, *83*(3), 1–21. https://doi.org/10.1007/s10670-017-9894-2
Sturgeon, S. (2008). Reason and the grain of belief. *Noûs*, *42*(1), 139–165. Retrieved from http://www.jstor.org/stable/25177157
Taroni, F., Bozza, S., Biedermann, A., & Aitken, C. (2015). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, *15*(1), 1–16. https://doi.org/10.1093/lpr/mgv008
Thompson, W. C., Taroni, F., & Aitken, C. G. G. (2003). How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Science*, *48*(1), 47–54.
Titelbaum, M. G. (2020). *Fundamentals of bayesian epistemology*.
Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman; Hall London.
Weatherson, B. (2002). Keynes, uncertainty and interest rates. *Cambridge Journal of Economics*, *26*, 47–62.