Bayes nets and the dynamics of probabilistic language

Daniel Lassiter — *Stanford University*

Abstract. This paper is about a shared concern of linguistic semantics & pragmatics, epistemology, and many other areas of cognitive science: the formal representation of information and uncertainty. It is common in many of these areas, and increasingly in linguistics, to represent agents' information using probability—an enrichment of classical semantics. However, each application of probability must answer difficult questions about whether a probabilistic representation is rich enough to capture the nuances of our information states. Two such problems—the epistemological distinction between uncertainty and ignorance, and the dynamic effects of probabilistic language in formal pragmatics—seem to suggest a negative answer, and to support a more complicated model that repesents uncertainty in terms of *sets of* probability measures. Following insights due to Judea Pearl (1988), I argue that the simplest probabilistic approach may be sufficient to handle these problems if we pay close attention to the *hierarchical* structure of information states, as encoded explicitly in the graphical representation of relationships among questions known as "Bayesian networks" or "Bayes nets".

Keywords: epistemic modality, probability, dynamics, epistemology

1. Three scenarios and two puzzles

(1) Two teams, A and B, are about to compete in a soccer game. You've seen them compete many times, and you are certain that they are evenly matched. What probability should you assign to the sentence "A will win"?

(2) Two teams, A and B, are about to compete at soccer. You know nothing at all about these two teams. What probability should you assign to the sentence "A will win"?

If pushed, most people will give the same answers to these questions: "50%". But our reason for giving these answers is obviously different in (1) and (2). In (1), we have a lot of relevant information to justify making this choice with confidence. In (2), our choice is made in ignorance: we just don't have any reason at all to favor one team over the other. Obviously, there is an epistemologically relevant difference, and it would be a mistake to represent our information identically in (1) and (2). But probabilistic representations seem not to make a distinction here.

(3) As in example (2), there are two teams, A and B, who are about to compete, and you know nothing at all about them. However, a knowledgeable friend assures you that "Team A is likely to win." What probability should you assign to "A will win"?

Here we move from a mainly epistemological question to one that lies at the border of epistemology, formal pragmatics, and computational cognitive science. Epistemologists and psychologists worry about the way that people do (or should) modify their states of information in light of *new information, however acquired*. Formal pragmaticists and dynamic semanticists worry about the way that people do (or should) modify their states of information in light of *new information acquired by linguistic means*. The latter kind of question being a special case of the former, both groups of theorists should care about what information is conveyed by the sincere assertion of a statement with probabilistic import, like “Team A is likely to win”. This is a particularly important problem for theorists of a Bayesian bent, since for them probability—or “credence”—is the basic currency of belief. Unfortunately, there is no general theory of the informational effects of probabilistic language to date. An attempt to frame one must balance three considerations, each difficult in its own right: getting the empirical facts right, integrating with existing accounts of linguistic dynamics, and remaining plausible in light of a broader accounts of epistemology and the cognitive science of learning and reasoning.

This paper treats the dynamics of probabilistic language and the confidence/ignorance distinction as two sides of an epistemological coin. The first approach takes ordinary probability to be inadequate as a representation of agents’ credence states, and opts for a richer model using imprecise credences—i.e., sets of probability measures. Probabilistic language, in turn, is modeled in terms of filtering on this set. The second approach tries to explain both the confidence/ignorance distinction and the dynamic effects of probabilistic language in terms of a single probability model that incorporates a hierarchical structure—such as that of a Bayes net. The need to include hierarchical structure in probabilistic models employs enjoys vast psychological, philosophical, and computational motivation (e.g., Pearl 1988, 2000; Glymour 2001; Sloman 2005; Woodward 2003; Tenenbaum et al. 2011; Danks 2014). Hierarchical models can also be used to model the interplay between statistical and causal reasoning, which is crucial in many linguistic, cognitive, and philosophical applications. There is an obvious gain in theoretical simplicity, then, if we can apply this independently motivated modeling approach to resolve the problems discussed here as well. However, my primary argument in favor of the hierarchical approach will be that it yields a better account of the basic empirical facts.

None of this calls directly into doubt whether further phenomena might motivate the use of imprecise probabilities in epistemology, psychology, or linguistics. Nor does it bear on the rather different question of whether imprecise-probability models are useful in modeling epistemic phenomena that extend beyond the minds of individuals, such as group belief or conversational common ground. (See brief comments on the latter in §5 below.) The main claim of the current work is rather that certain phenomena which appear to problematize Bayesian models of individual agents’ informational states and their (linguistic and non-linguistic) dynamics in fact already have an illuminating explanation within these models.

2. Precise and imprecise credences

Say that an agent a has *precise credences* just in case a’s state of information is well-represented by a probability measure P_a. Since this function is being used here to model a’s state of belief, or “credence”, we’ll also call it a “credence function”. P_a assigns a number between 0 and 1 to each proposition $A \in \wp(W)$, subject to the usual laws of probability (Kolmogorov, 1933): $P_a(W) = 1$, and $P_a(A \cup B) = P_a(A) + P_a(B) - P_a(A \cap B)$. (These definitions assume that W is finite, inessentially, in order to simplify the math.) Just what it means for a’s information to be “well-represented” by P_a is a difficult theoretical question that I will leave at an intuitive level here. Explicit judgments about probability, other linguistic behaviors, overt choices, and dispositions to choose in particular ways are some of the many ways that we might want to evaluate whether a candidate P_a is a good representation of a's information. Notice that this model does not compete with classical semantics, but rather *presupposes* and *extends* it: a set of worlds W generates a Boolean algebra $\wp(W)$ whose elements are propositions that receive probabilities. Probability assignments are constrained by classical logic. For example, if A and B are contradictory—$A \cap B = \varnothing$—then $P(A \cap B) = 0$. If A entails B—$A \subseteq B$—then $P_a(A) \leqslant P_a(B)$. (Indeed, the probabilistic model inherits some important limitations of the classical semantics, such as problems around logical omniscience and hyperintensionality.)

Precise credence models have many epistemological and cognitive advantages, and are also subject to many kinds of objections. One well-known objection involves experimental evidence that ordinary people make systematic errors in probabilistic reasoning (e.g., Tversky and Kahneman 1974; Kahneman et al. 1982). While this kind of critique is surely relevant, I want to set it aside here with a few quick comments. First, there are many additional experiments in which people seem to reason appropriately with probabilities. Second, experiments in which people are asked to reason explicitly about probabilities may be less theoretically revealing than those in which probabilistic reasoning is implicit in the way that uncertainty informs judgment and decision (e.g., Griffiths and Tenenbaum 2006; Trommershäuser et al. 2008). The logic is essentially the same as that which motivates linguists to pay closer attention to unreflective linguistic productions and judgments than to explicit metalinguistic judgments. Third, recent work has suggested a reconciliation, where at least some errors and biases in probabilistic reasoning may be explicable in terms of performance factors, interactions among cognitive systems, or strategies for efficient approximation (Griffiths et al., 2012; Vul et al., 2014).

The objections that motivate imprecise credence models are primarily of a different kind. Kahneman & Tversky argued that ordinary people's credence states, to the extent that they are not consistent with a precise credence model, fail to meet a normatively correct epistemological standard. Proponents of imprecise credences, in contrast, argue that it would in many cases be normatively inappropriate for an agent to have a credence state that is consistent with a precise credence model. A typical example is scenario (2): two teams compete in a game, and you know nothing at all about their relative skills. Joyce (2005, 2010) argues that, in such a scenario, you are making a mistake if you have *any* precise credence in team A winning. What possible grounds could you have for such "extremely definite beliefs ... and very specific inductive policies", when "the evidence comes nowhere close to warranting such beliefs and policies" (Joyce 2010, p.285)? Depending on the teams' relative skills, the right credence to have might be any value in the range $[0,1]$! You don't know enough to exclude *any* of these.

This objection is closely related to the problem of insufficient expressiveness that we began with. When asked for a probability estimate in scenarios (1) and (2), I might produce "50%" in both cases—but confidently in (1), and with hesitation and confusion in (2). Similarly, I would immediately reject an uneven bet on either team in (1), but might have a harder time making up my mind in (2). Either way, the precise model seems to miss at least two important differences between these judgments: differences in their evidential basis, and in their phenomenology. Any two events to which I give credence 0.5 just have credence 0.5, end of story. As a result, precise credences are not fine-grained enough to provide a good model of my credence state. As Halpern (2003, p.24) puts it, "Probability is not good at representing ignorance".

3. Confidence and ignorance: An imprecise model

The proposed alternative is to represent an agent a's information not by a single measure P_a, but by a *set* of measures $\mathbb{P}_a$ (e.g., Levi 1974; Jeffrey 1983). This is sometimes called a's "Representor" (van Fraassen, 1990). Each $P \in \mathbb{P}_a$ conforms to the probability axioms. This model has no expressive difficulty in the sporting examples. In the first scenario, where I am confident that the teams are evenly matched, my $\mathbb{P}$ has the property that, for every $P \in \mathbb{P}$, $P(\text{A wins}) = 0.5$. In the second scenario, where I have no relevant information, my $\mathbb{P}$ has the property that, for every $r \in [0,1]$, there is a measure $P \in \mathbb{P}$ such that $P(\text{A wins}) = r$. In the first case I have an "extremely definite belief" (Joyce, 2010) that $P(\text{A wins}) = 0.5$, and I am right to. In the second I have no definite belief about the value of $P(\text{A wins})$, and I am right not to.

Despite this apparent success, some important objections have been made to the use of imprecise probabilities. One is that it is difficult to frame a plausible decision theory for agents with imprecise credences. Elga (2010), in particular, canvasses a number of options and shows that each makes pathological predictions in certain cases; see also White 2010. A second kind of objection involve examples where imprecise models seem to predict, rather oddly, that learning a new fact should lead to a net loss of information ("probabilistic dilation": Seidenfeld and Wasserman 1993), or where learning something that is intuitively irrelevant to an event A leads to a gain of information about A's probability (White, 2010). (However, see Pedersen and Wheeler 2014 for important subtleties that may help to improve the plausibility of dilation.) These particular objections are two of many, and they are still a matter of active controversy in the epistemological literature. I don't want to take a stand on whether they are decisive, but I do think they give us sufficient reason to look for a model that combines naturally with well-understood, well-behaved Bayesian models of learning and decision. First I will discuss a third objection which introduces some of the motivation for the hierarchical alternative.

Perhaps the most troubling objection to imprecise probability models, from our current perspective, is the observation that they "preclude[] inductive learning in situations of extreme ignorance" (Joyce 2010, p.290; see also White 2010; Rinard 2013). For example, consider a biased coin example analogous to scenario 2 above. Suppose I am uncertain about the probability of getting heads when a certain coin is tossed. This probability could in principle be anywhere in $[0,1]$. On any given toss, the probability of getting heads—$P(\text{heads})$—is equal to the coin's bias π, which is a fixed fact about the world, determined by the coin's objective properties. My uncertainty about $P(\text{heads})$ reduces to uncertainty about the value of π.

Suppose I had precise credences, with a prior distribution on π—say, a Beta distribution. If I wanted to be maximally noncommittal, I might use a $\text{Beta}(1,1)$ distribution, which puts equal prior probability on every bias $\pi \in [0,1]$ (see Fig. 1, left). Given this model, after conditioning on the observation of n heads and m tails my posterior probability is given by $\text{Beta}(1+n, 1+m)$. (In general, conditioning a $\text{Beta}(a,b)$ prior on n heads/successes/wins and m tails/failures/losses yields a $\text{Beta}(a+n, b+m)$ posterior. See Griffiths et al. 2008; Hoff 2009 for introductions to the Beta-Binomial model.) So, for example, if I had a maximally noncommittal $\text{Beta}(1,1)$ prior, after observing 150 heads in 300 tosses my beliefs about the bias π would be updated to a $\text{Beta}(151,151)$ distribution. This prior-to-posterior mapping is pictured in Fig. 1. The quite

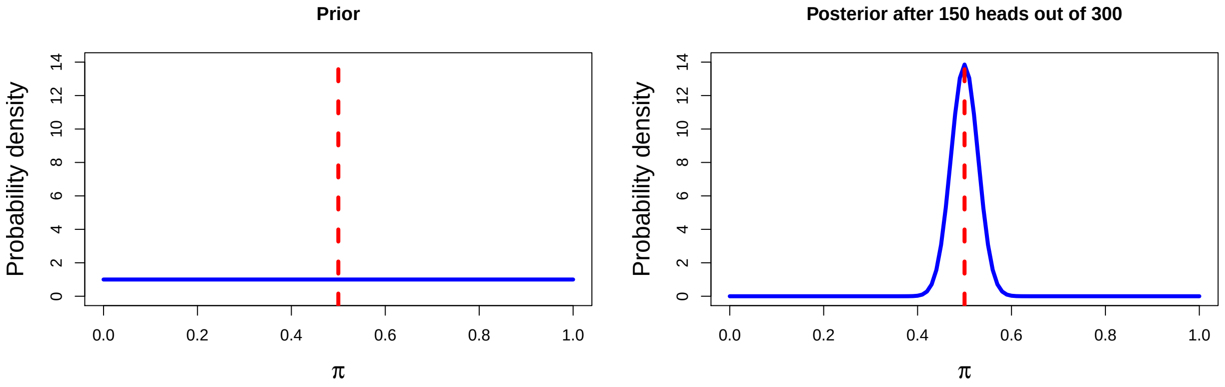


Figure 1: Prior-to-posterior mapping for an agent with precise credences and a Beta(1,1) prior, after observing 150 heads/successes out of 300 trials. The red line indicates the expected value of the parameter π, which does not change with this evidence even though our uncertainty about the estimate (i.e., the variance of π) decreases dramatically.

reasonable prediction is that, after observing 150/300 heads, you can be quite confident that the coin's bias π is close to 0.5—even if you were maximally noncommittal about π to begin with.

Not so in the imprecise credence model with the standard update rule of pointwise conditioning. I have no idea about the probability of heads initially, so my initial representor $\mathbb{P}_0$ contains, for every $r \in [0,1]$, a credence function P such that $P(\text{heads}) = r$. For example, $\mathbb{P}_0$ might contain, for every possible Beta prior, a measure that encodes a binomial model with that prior.

$$\mathbb{P}_0 = \{P \mid P(\pi) \sim \text{Beta}(a,b),\ \ \forall a,b \in [0,\infty)\}$$

(Using only Beta priors is a significant restriction relative to Joyce's philosophical desiderata, but using the full range of possible distributions would only make the problem worse.) Now, suppose I observe 150/300 heads and update $\mathbb{P}_0$ to $\mathbb{P}_1$ by pointwise conditionalization, discarding measures that assign probability 0 to the observations and so cannot generate the sequence. In this case, the latter condition requires us to discard any Beta prior with a 0 in either position, which could only generate "all heads" or "all tails" sequences. All other measures in $\mathbb{P}_0$ assign positive probability to the observed sequence of 150 heads and 150 tails, and survive in conditionalized form as $\text{Beta}(a+150, b+150)$ measures:

$$\begin{aligned} \mathbb{P}_1 &= \{P \mid P(\pi) \sim \text{Beta}(a+150, b+150),\ \ \forall a,b \in (0,\infty)\} \\ &= \{P \mid P(\pi) \sim \text{Beta}(a',b'),\ \ \forall a',b' \in (150,\infty)\} \end{aligned}$$

When we look at a few of these distributions, it is clear that something has gone wrong. Alongside reasonable-ish posteriors like $\text{Beta}(160,200)$ [so $P(\text{heads}) \approx .44$] and $\text{Beta}(200,160)$ [so $P(\text{heads}) \approx .56$], the posterior belief state contains a $\text{Beta}(150.1, 10^{14})$ posterior [where $P(\text{heads}) < 10^{-10}$] and a $\text{Beta}(10^{14}, 150.1)$ posterior, where $P(\text{heads})$ is indistinguishable from 1. This is truly remarkable, since the probability that we would have seen 150 or more heads in 300 if $P(\text{heads}) = 10^{-10}$ is around 10^{-87}—but this failure of prediction is not taken into account in update by pointwise conditioning. In fact, for every r in the open interval $(0,1)$, there is a measure in $\mathbb{P}_1$ such that $P(\text{heads}) = r$. As far as the spread of probabilities for heads is concerned, all that we have gained from our observations is to contract the interval $[0,1]$ to $(0,1)$, ensuring that both heads and tails are *possible* outcomes. We have learned *nothing else* about the coin's bias. But in reality, a sequence of 150 heads and 150 tails can and should teach us a lot, even if we know nothing at all about the coin to begin: it is almost certainly fair, or very close to it. Inductive learning *is* possible from a starting point of ignorance.

Several responses are possible here. First, we could search for an alternative to pointwise conditioning as an update rule. I won't speculate about how this would go. A second option would be to rule out representors where $P(\text{heads})$ may fall anywhere in $[0,1]$ or $(0,1)$. This would avoid the narrow problem addressed here, but it seems poorly motivated. If imprecise credences are motivated in the first place by considering belief under ignorance, how can we justify dealing with theoretical problems by *pretending* to know something that we don't? (Never mind that many such restricted models will still exhibit no learning, or will learn at an unbelievably slow rate.) A third option, floated by Joyce (2010), is to conclude that it is in fact *not* possible to learn in a rational way from a starting point of total ignorance. However, Joyce continues, real people employ non-rational heuristics to help them get by psychologically, such as restricting attention to measures that give high enough probability to the observed evidence. This response seems desperate, and if taken seriously it may imply that all of our beliefs are irrationally held: after all, for each of my beliefs there was some point at which I was totally ignorant on the subject.[1] A fourth option is to take this problem to demonstrate the impossibility of providing a precise formal model of belief states (Rinard, 2013). This could be correct, but I hope it isn't.

My preferred response is to reject imprecise credences as a model of individuals' belief states. To plump for this option, let me point out the key technical difference between precise- and imprecise-credence models that is creating this problem: whether we place a probability distribution on top of the set of credence functions in $\mathbb{P}_0$. Imprecise models decline to assign probabilities to the elements of $\mathbb{P}_0$, leaving it as an unstructured set. If we did put a distribution on $\mathbb{P}_0$, we would end up with a precise-credence model with a hierarchical structure, as I will describe in the next section. In this case, many kinds of (hyper-)priors on $\mathbb{P}_0$ would yield plausible results with ordinary conditioning. We can see why this small change makes a difference if we break down conditioning using Bayes' rule. With a distribution on the measures in $\mathbb{P}_0$, the posterior probability of each $P \in \mathbb{P}_0$ would be proportional to the product of the prior and the likelihood, where the latter is the probability that we would have observed the data if P were the true distribution. Conditioning re-ranks credence functions to take into account such facts—e.g., that 150/300 heads is moderately likely under a $\text{Beta}(160,200)$ distribution, and astronomically unlikely under a $\text{Beta}(150.1,10^{14})$ distribution. In contrast, imprecise models do not represent information about the relative plausibility of the measures in the representor, and pointwise conditioning does not take into account how well the measures in $\mathbb{P}_0$ fare in the goal of predicting the data (a likelihood term). This seems to be the basic reason why imprecise models fare so poorly when confronted with simple examples of inductive learning.

In order to extract a plausible treatment of learning from imprecise credence models we need

[1] We all began, in the womb, in a state of total ignorance, though we must presumably have been endowed with inductive biases. As psychologists and machine learning researchers are fond of reminding us, learning without initial biases (i.e., a hypothesis space and priors) is impossible: see, for example, Perfors 2012.

to put a distribution on $\mathbb{P}_0$ so that we can apply ordinary conditioning. In other words, we need a prior on our priors, which is the basic idea of hierarchical models.

4. Confidence and ignorance: A hierarchical model

This is, to be sure, a roundabout way of getting to a simple objection. We just don't *need* imprecise credences to represent the difference between confidence and ignorance in the sporting scenarios we started with. Arguments against precise models based on a supposed failure to represent this distinction are simply misdirected. Confidence and ignorance can be given a satisfying treatment within hierarchically structured models, which are well-developed formally and strongly motivated psychologically and computationally.

Recall Joyce's (2010, p.285) objection to precise models quoted above: in a situation of ignorance, it is not justifiable for you to have "extremely definite beliefs ... and very specific inductive policies", because "the evidence comes nowhere close to warranting such beliefs and policies". Already in the coin-bias example, though, this objection is at least partly misplaced.[2] If your prior on the bias parameter π is a $\text{Beta}(1,1)$ distribution (Fig. 1, left), your belief is anything but definite. It is true that π has a precise expected value 0.5, and also that your marginal belief about $P(\textbf{heads})$ is 0.5. However, you are extremely uncertain about both of these beliefs: depending on what evidence you receive, you could come to a very definite conclusion that π and $P(\textbf{heads})$ are both 0, both 1, or anywhere in between. For example, after observing 0/300 heads, your posterior distribution on π would be $\text{Beta}(1,301)$, with $P(\textbf{heads})$ indistinguishable from 0. *This* would be an "extremely definite" opinion, and one that is justified by the evidence. Similarly, after seeing 150/300 heads, you have a fairly definite opinion that π and $P(\textbf{heads})$ are close to 0.5 (Fig. 1, right). Even though the summary estimate $P(\textbf{heads}) = 0.5$ (red dashed line) does not change, the transition from the information state described by the left of Fig. 1 to the one on the right clearly represents a significant change in your beliefs about $P(\textbf{heads})$.

More generally, I suggest—building on observations made in another context by de Finetti 1977 and Pearl 1988 (p.357ff.)—that many of the intuitive arguments for imprecise probabilities discussed above can be accounted for in a better-motivated way once we take into account the hierarchical structure of belief. Our beliefs are interconnected, and probability estimates involving one variable usually depend on uncertain beliefs about others. Uncertainty about one variable—e.g., the bias π of a coin—may influence our uncertainty about a probability estimate of interest—e.g., the probability that the coin will come up heads on a given flip. Given the richness of our belief systems, there will usually be many layers of uncertainty. Even though such a model will always yield a precise numerical probability for any event of interest, this

[2] The part that may well hit home nonetheless is the accusation that precise credence models give rise to "very specific inductive policies" which are not warranted by the evidence. This is essentially the same point as the correct observation that precise credence models require priors that are not chosen on the basis of experience. Some authors have argued that priors should be chosen so as to maximize entropy (Jaynes, 2003; Williamson, 2009), though there are well-known objections to this move (van Fraassen, 1989). Does this mean that no choice of priors can be uniquely justified? Perhaps, but it is not clear why this should be so troubling. Since learning without priors is impossible, Bayesian cognitive models imply that evolution has supplied us with priors that are good enough to enable successful learning, starting in the womb. These may well vary between individuals, and there is no reason to expect that nature's choice of priors should be uniquely justified or rational. All we can expect is that they should get the job done with respect to the ultimate goals of survival and reproduction.

numerical value does not have any special place in the model: it is just what you get when you marginalize over your uncertainty about other relevant variables. In a hierarchical model, probability estimates can vary enormously in how "definite" they are—more precisely, in their variance.

Hierarchical models are used in many modern applications in psychology, philosophy, artificial intelligence, and statistics. In these models, probabilities are derived from graphs representing causal relations among variables, together with the conditional distribution on each variable given its parents. Uncertainty about one variable may influence the kind and degree of uncertainty in the value of another. For simplicity I will focus on Bayesian networks ("Bayes nets"), a simple propositional language for describing hierarchical models. (For discussion of richer languages for describing hierarchical Bayesian models that can treat uncertainty about individuals, properties, relations, etc., see for example Goodman et al. 2008; Goodman and Tenenbaum electronic; Tenenbaum et al. 2011 and, for a linguistically-oriented presentation, Goodman and Lassiter 2015.) I will impose a causal interpretation on the Bayes nets described in this paper. While this is not obligatory, it helps to gain intuitions about their meaning, and it is crucial to their psychological motivation (e.g., Glymour 2001; Sloman 2005).

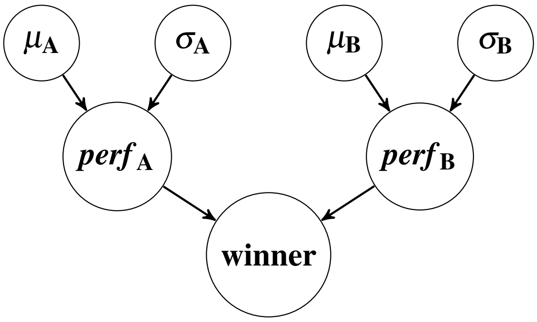
The sporting example that we began with allows us to illustrate Bayes nets and their ability to represent confidence and ignorance alike. Formally, a Bayes net B is an enrichment of familiar intensional semantics models, consisting of a set of possible worlds W together with:

1. A set of "variables" $V \in \mathbb{V}$, where each V is a partition of W. (A cell is a "value" of V.)
2. A set of "arrows", i.e., an acyclic binary relation on $\mathbb{V}$. (The inclusion of an arrow from V_i to V_j indicates that V_i is immediately causally relevant to V_j.)
3. A set of conditional probability tables which assign a distribution $P(V \mid Parents(V))$ to each $V \in \mathbb{V}$, where $Parents(V) = \{V' \mid \langle V', V\rangle \in \mathbb{V}\}$.

A probability measure P is *compatible* with Bayes net B if and only if P satisfies the *Markov condition*: each $V \in \mathbb{V}$ is probabilistically independent of its nondescendents, given its parents.

To situate the hierarchical modeling concept within our sporting examples, consider: In case (2), when asked to reason about the competition between unknown teams **A** and **B**, did you really know *nothing at all* about these teams? I doubt it. Most likely, you brought to bear on the problem a rich network of relevant background knowledge. You knew that the outcomes of matches are determined largely by the performance of the teams; that teams are composed of players who have different roles; that they have latent characteristics like skill and consistency; that not all teams are equally skilled or consistent; and so forth. In addition, your experience may have provided you with relevant population statistics which can help you to make an informed guess about the distribution of these characteristics among teams, even without any specific knowledge of the team. All of this background knowledge enabled you to make a reasonable guess about how a randomly chosen team would perform, and what factors you should attend to if you want to use observations to improve your forecast of a team's performance.

As a start in modeling the richer background knowledge that we implicitly bring to bear on such problems, consider the simplified representation in Fig.2.[3] This model represents two key features of teams that are relevant to their performance: their **skill** and their **consistency**. Performance of team i is modeled as a Gaussian (normal) distribution with parameters μ_i (**skill**) and σ_i (**consistency**). As a result, the team's performance in any given competition is a noisy reflection of the team's true skill. Skill and consistency are, in turn, objects of uncertainty that we are trying to estimate when observing the outcomes of competitions. This means that we must place a prior on them as well. In a realistic model, these variables might be connected to many further factors—e.g., the team's composition, quality of coaching, motivation, etc. To simplify the example, I will summarize all of these sources of uncertainty with simple priors on the parameters: $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$ are both distributed as $\mathcal{N}(0,1)$, and $\sigma_{\mathbf{A}} = \sigma_{\mathbf{B}} = .1$.



$$\forall i : \mu_i \sim \mathcal{N}(0,1)$$
$$\forall i : \sigma_i = .1$$

$$\forall i : perf_i \sim \mathcal{N}(\mu_i, \sigma_i)$$

$$\textbf{winner} = \mathbf{A} \text{ iff } perf_{\mathbf{A}} > perf_{\mathbf{B}}$$

Figure 2: Hierarchical model of a match between teams **A** and **B**.

In this model $P(\textbf{A wins})$ is equal to $P(perf_{\mathbf{A}} > perf_{\mathbf{B}})$—the probability that **A**'s noisy performance exceeds **B**'s. Note that this model does *not* generate a single, determinate prediction about **A**'s performance in any given match. Instead it generates for each team a distribution over an infinite set of performance values $(-\infty, \infty)$. A few of these distributions are shown in the top left of Fig. 3. As a result, the model encodes a distribution over an infinite set of values for $P(\textbf{A wins})$, which could be anywhere in $(0,1)$ depending on subsequent observations.

While the model does yield a precise best guess about the performance difference—and so about $P(\textbf{A wins})$—this guess has no special status in the model: it is merely the result of marginalizing over our uncertainty about the parent variables (skill and consistency). Indeed, two models that generate the same probability estimate for this event—say, $P(\textbf{A wins}) = .5$—may vary considerably in how confident ("definite", "determinate") the probability estimate is.[4] A key factor is, of course, how much evidence the estimate is based on.

[3] The model is directly inspired by the Microsoft Trueskill system that is used to rank Xbox Live players in order to ensure engaging match-ups in online games: see Bishop 2013. It is also closely related to the tug-of-war model explored by Gerstenberg and Goodman (2012); Goodman and Lassiter (2015).

[4] As Pearl (1988, p.361-2) writes: "The point is to notice that by specifying a causal model for predicting the outcome ... we automatically specified the variance of that prediction. In other words, when humans encode probabilistic knowledge as a causal model of interacting variables, they automatically specify not only the marginal and joint distributions of the variables in the system, but also a particular procedure by which each marginal is to be computed, which in turn determines how these marginals may vary in the future. It is this implicit dynamic that makes probabilistic statements random events, admitting distributions, intervals, and other confidence measures." As a consequence, Joyce (2010, p. 283) is simply wrong in his assertion that "Proponents of precise models ... all agree that a rational believer must take a definite stand by having a sharp degree of belief" in situations of ignorance. While Joyce is right that taking a "definite stand" in case of ignorance is unreasonable, his conclusion

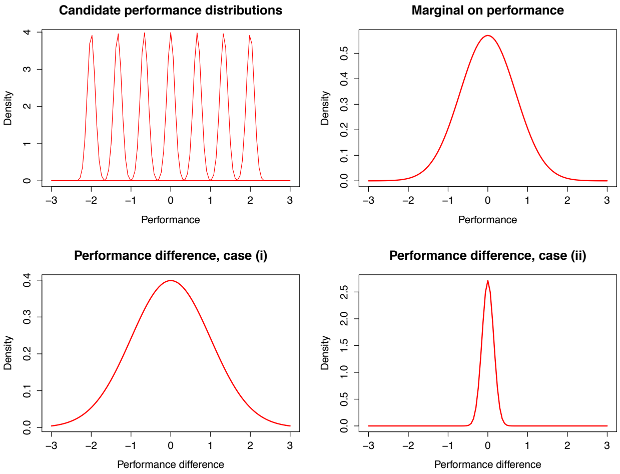


Figure 3: Some distributions implicit in the Fig. 2 model. Top left: some of the ∞ performance distributions that could turn out to be the true distribution for either team. Top right: Marginal on $\mathbf{perf}_{A/B}$ with no specific evidence. Bottom row: Distribution of $\mathbf{perf}_A - \mathbf{perf}_B$ with no observations (left) and after observing that each team won 15 of 30 matches (right).

Consider our two leading examples again. In case (1), we "know nothing"—i.e., only general domain knowledge is available. As a result, the variance of the estimated performance difference is high, and confidence in the estimate $P(\mathbf{A\ wins}) = .5$ is low (Fig. 3, bottom left). In case (2), there is ample evidence to indicate equal skill—many previous matches, with each team winning an equal number. In this case, the variance of the estimated performance difference is low, and confidence in the estimate $P(\mathbf{A\ wins}) = .5$ is high. The bottom right panel of Fig. 3 shows the model's predictions about $P(\mathit{perf}_{\mathbf{A}} > \mathit{perf}_{\mathbf{B}})$ once we have observed each team winning 15 of 30 matches. Here we can infer that the teams have roughly equal skill, and that we should forecast roughly equal performance in the next game: $P(\mathbf{A\ wins}) = .5$.

Bayes nets offer a precise credence model that represents the distinction between confidence and ignorance in a straightforward way. The need to represent this distinction does not, therefore, give us a reason to abandon precise credence models in favor of a more complex representation that also introduces new problems involving learning and decision. The apparent problem with precise models—that radically different credal states could generate the same probability estimate $P(\mathbf{A\ wins}) = .5$—was not due to any expressive limitation. Instead, the problem is generated by our habit as theorists of treating such numerical estimates as representations of belief states, forgetting that they give only a narrow window into the rich structure of a probability distribution and its potential to change in response to experience.

5. Dynamics of probabilistic language: An imprecise model

The third puzzle that we started with was how to understand the informational effect of explicitly probabilistic language. Example (3) is repeated here:

that precise credence models are inappropriate is a non sequitur: precise probability estimates can be extracted from hierarchical models, but these models do not generally imply a "definite stand" on these estimates.

(4) Two teams A and B who are about to compete, and you know nothing at all about them. However, a knowledgeable friend assures you that "Team A is likely to win." What probability should you assign to "A will win"?

This puzzle is a special case of the more general question of how we should update our beliefs on the basis of new information. However, in (3) the new information is provided in linguistic form, and so we have the dual problem of supplying an interpretation and an update procedure that will generate the right mapping from information states to modified information states.

In familiar models of linguistic dynamics, information states I are usually modeled as sets of worlds—or sometimes as sets of structured objects that segment out different kinds of information and handle them separately. Sentences S pick out, relative to a context c, properties of the kind of objects contained in I. So, if I is a set of worlds, $[\![S]\!]^c$ is a set of worlds as well. The algorithm for incorporating the information conveyed by S in c into state I is simple: we transition from state I to state $I \cap [\![S]\!]^c$, eliminating from I any worlds that are not in $[\![S]\!]^c$. On this approach to conversational dynamics, information states and sentence-meanings-in-context must be objects of the same type. This means that, if our language contains sentences whose meaning is not well treated as a set of worlds, an eliminative theory of update requires us to enrich the representation of information states accordingly.[5]

Now, recent work has given considerable semantic motivation to the claim that the interpretation of *A is likely to win* in (4) makes direct reference to a probability—roughly, "$P(\textbf{A wins}) > .5$" (Swanson, 2006; Yalcin, 2007, 2010; Lassiter, 2010, 2011, 2017a; Moss, 2015). This leaves us with a theoretical dilemma. If we want to hold to the eliminative conception of update, we can either try to give this probability statement a world-relative interpretation, or we can assign it some other interpretation and enrich our model of information states to compensate.

Yalcin 2012 proposes a simple eliminative model of update for both factual and probabilistic statements, building on Yalcin 2007. On this account epistemic statements—including probabilistic statements—do not have a world-relative interpretation. Instead, they depend on the value of a *sui generis* information state parameter which varies independently of the choice of evaluation world, and which determines a probability measure P. Briefly, the idea is that a conversational common ground C is a set of pairs $i = \langle s_i, P_i \rangle$, where s_i is a set of worlds and P_i is a probability measure with $P_i(s_i) = 1$. For factual statements such as *It's raining*, update eliminates from C those i for which it is not raining everywhere in s_i, without constraining P_i.

$$I \underset{\text{update}}{\Longrightarrow} I \cap \{\langle s, P \rangle \mid \forall w \in S : \text{It's raining at w}\}.$$

For probabilistic statements such as *It's likely to rain*, update eliminates points $i \in C$ that do not assign sufficiently high probability to rain, but places no direct constraints on s_i.

$$I \underset{\text{update}}{\Longrightarrow} I \cap \{\langle s, P \rangle \mid P(\textbf{rain}) > .5\}.$$

[5] An alternative is to modify the definition of update, giving pointwise definitions for individual expressions, as Veltman (1996) does for *might* and several other epistemic operators. However, this approach is not promising for *likely*, and I will not consider it further.

This formal model of conversational dynamics bears a striking resemblance to the sets-of-measures model of credence that we considered above. However, there are two very different ways to interpret it. Yalcin's proposal is to use it to model the dynamics of conversational common ground, in the mold of Stalnaker 1978. The idea is that, as a conversation proceeds, the acceptance of utterances leads to an accumulation of constraints on interlocutors' information. Once "It's likely to rain" is accepted, the common ground contains the information that $P(\mathbf{rain}) > .5$, meaning that interlocutors are publicly committed to this constraint. This interpretation is conceptually close to the "group belief" interpretation of imprecise probability models discussed above. As mentioned above, I have no quarrel with the group belief interpretation of imprecise credences, where measures in the set represent assignments of probability that are consistent with existing group commitments; the objections canvassed were specifically leveled at an application to individual psychology. For similar reasons, Yalcin's interpretation seems unproblematic as a way of treating the way that conversational commitments constrain the way that common ground constrains (but does not determine) probability assignments.

Solving the puzzle in (3)/(4) requires a more ambitious theory, though. It is not enough to require that someone who accepts *A is likely to win* must assign $P(\mathbf{A\ wins}) > .5$. This constraint does not tell us enough about what that person *should* come to believe about $P(\mathbf{A\ wins})$, other than the bare fact that $P(\mathbf{A\ wins})$ cannot fail to exceed .5. What we are interested in is a mapping from an *individual*'s prior state of information to a posterior state, which would determine what she should believe if she begins in state I and then gains some explicit information about the likelihood of events. While it was not proposed for this purpose, Yalcin's formal apparatus could equally be put to use in this way (cf. also Rothschild 2012). On this interpretation, an individual's state of information I is a set of pairs $i = \langle s_i, P_i \rangle$, and we have a simple algorithm for updating with the information conveyed by *A is likely to win*: as in Yalcin's proposal, map I to the subset of I containing all $\langle s_i, P_i \rangle$ for which $P_i(\mathbf{A\ wins}) > .5$.

While this proposal is *prima facie* plausible, it encounters some serious difficulties. First, as a representation of individual psychological states its plausibility is threatened by general objections to the use of imprecise credence models for this purpose: problems with framing a decision theory, with updating beliefs from a starting point of ignorance, and so forth. Second, even if we were to accept imprecise credence models along the lines of Joyce (2005, 2010), the update procedure just described would not be sufficiently general. The problem is that imprecise probabilities are intended to represent situations where probabilities are not known—but probabilistic statements can be informative about known and unknown probabilities alike.

Consider a scenario in which a ball is drawn from one of two urns. Urn A contains 10 red and 5 blue balls, so $P(\mathbf{red} \mid \mathbf{A}) = 2/3$. Urn B contains 5 red and 10 blue balls, so $P(\mathbf{red} \mid \mathbf{B}) = 1/3$. A fair coin was flipped to determine which urn the ball would come from—A if heads, B if tails. So, $P(\mathbf{A}) = P(\mathbf{B}) = 1/2$. If someone who knows what urn was selected tells us *The ball is probably red*, how should we respond? Clearly, we can conclude that the coin toss came up heads and A was selected. So the posterior probability of **red**, after update with *The ball is probably red*, should be equal to $P(\mathbf{red} \mid \mathbf{A}) = 2/3$.

However, it is not possible to model this kind of update by filtering in a standard imprecise probability model: paradoxically, we know too much. That is, we could attempt to represent this situation in terms of an imprecise model where some measures have $P(\textbf{red}) = 2/3$ and some have $P(\textbf{red}) = 1/3$. Then, upon learning that the ball is probably red, we could model the update effect of this information by eliminating the measures where $P(\textbf{red}) = 1/3$. But as Joyce (2005, 2010) emphasizes, this model is inappropriate: it leaves out known probabilities, i.e., the information that the choice of urn was determined by the flip of a fair coin. Since $P(\textbf{A}) = P(\textbf{B}) = 1/2$, all measures P in our information state have

$$\begin{aligned} P(\textbf{red}) &= P(\textbf{red} \mid \textbf{A}) \times P(\textbf{A}) + P(\textbf{red} \mid \textbf{B}) \times P(\textbf{B}) \\ &= 2/3 \times 1/2 + 1/3 \times 1/2 \\ &= .5 \end{aligned}$$

But if $P(\textbf{red}) = .5$ according to all measures in our information state, the model of probabilistic update under consideration—"throw out all measures on which $P(\textbf{red}) \nsucc .5$"—will yield the empty information state. If only we didn't know that the choice of urn was determined by a fair coin toss, we could update with the information that the ball is probably red! The filtering approach to update with probabilistic language does not, it appears, play nicely with the kinds of imprecise credence models that have been given independent motivation in the philosophical and statistical literature.

6. Dynamics of probabilistic language: A hierarchical model

Our problem is to explain why, in the urn model just given, update with *The ball is probably red* leads to a coherent result—and to one which somehow yields the judgment that $P(\textbf{red})$ is precisely $2/3$. As in the general discussion of precise vs. imprecise credence models, I suggest that the solution to our problem is to pay closer attention to the hierarchical structure of information states. Consider a causal Bayes net representing this process (Fig. 4). The key thing to observe about this model is that the intuitive update effect of *The ball is probably red* is exactly the same as the intuitive update of *The coin came up heads*. Our goal is to find a way to use the structure of the Bayes net to guarantee that this will be the result, against this informational background. Once this is done, we may be able to define a general update procedure that applies to probabilistic and non-probabilistic language alike.

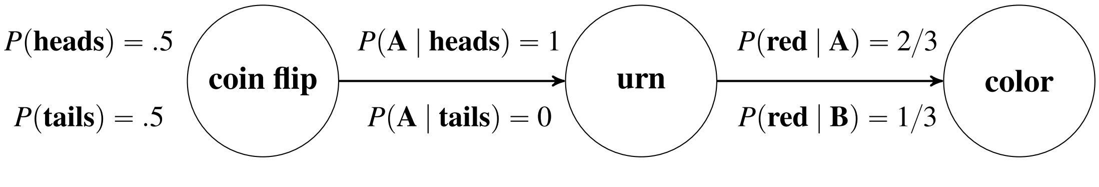

$P(\textbf{heads}) = .5$
$P(\textbf{tails}) = .5$
coin flip
$P(\textbf{A} \mid \textbf{heads}) = 1$
$P(\textbf{A} \mid \textbf{tails}) = 0$
urn
$P(\textbf{red} \mid \textbf{A}) = 2/3$
$P(\textbf{red} \mid \textbf{B}) = 1/3$
color

Figure 4: Causal Bayes net modeling the urn scenario.

A Bayes net defines a way of partitioning the space of possible worlds into increasingly fine-grained cells as a causal process unfolds. The first step in the process is the flip of a coin, which has two possible outcomes with specified probabilities. If we happen to be at a world w in which the coin came up heads, then an urn is selected. In this model, the urn is A with probability 1 in this case. So, w—and any other heads-world—is also an A-world. The red/blue mix of the urn selected in w is 10/5, and so the probability that the ball selected will be red is 2/3. The next step is the choice of a ball: at any w, either the ball chosen is actually red, or it is actually blue. Nevertheless, it is a fact about w that the probability of a red ball being selected is 2/3, even if the random selection process actually unfolds so that a blue ball is chosen. The Bayes net thus suggests a world-relative probability concept, where probabilities are dissociated from the actual occurrence of events. As long as our observations do not include the ball's color or any downstream effects of this variable, it is true at any such w that the ball is probably red, because this is a world at which the coin came up heads and the urn selected was A.

My suggestion, then, is to interpret probability statements as involving a kind of world-bound probability, but relativized to a Bayes net B. (At least probability statements with an "objective" flavor—see caveats below.) Suppose that the statement is *The ball is probably red*. Rather than looking for a global, world-independent parameter P and checking the value of $P(\mathbf{red})$ (as in Yalcin 2007, 2012), we look to $P_w(\mathbf{red})$, the probability that B assigns to **red** at world w. With a world-bound probability concept available, we could then model update with *The ball is probably red* as simple conditionalization, a procedure that works equally for factual statements. Here, update means conditioning on the set of worlds w' such that $P_{w'}(\mathbf{red}) > .5$.

(5) a. $[\![\text{The ball is probably red}]\!]^{B,w} = 1 \text{ iff } P_w(\mathbf{red}) > .5.$
    b. $P \underset{\text{update with (5a)}}{\Longrightarrow} P(\cdot \mid \{w' \mid P_{w'}(\mathbf{red}) > .5\}).$

If this is to work, the key question is how B and w conspire to determine P_w. Recall from above that the intuitive update resulting from *The ball is probably red* is the same as the update from *Urn A was selected*: both tell us, in effect, that $P(\mathbf{red}) = 2/3$. Relative to the Bayes net in Fig. 4, then, the set of worlds in which *The ball is probably red* is true should be the same as the set of worlds in which *Urn A is selected* is true. As a first approximation, we could define $P_w(\mathbf{red})$, relative to B, to be equal to the conditional probability of **red** in B given the actual value(s) at w of the immediate parent(s) of the variable of which **red** is a value. Since the only parent of **color** is **urn**, this means that $P_w(\mathbf{red}) = P(\mathbf{red} \mid \mathbf{urn}_w)$, where $\mathbf{urn}_w$ is the actual value of **urn** at w. If the urn is A in w, then $P_w(\mathbf{red}) = 2/3$. If the urn is B at w, then $P_w(\mathbf{red}) = 1/3$.

If we define P_w in terms of the structure of a Bayes net in this way, the contextual equivalence of *The ball is probably red* and *Urn A was selected* follows immediately. Relative to the Bayes net in Fig. 4, the set of worlds w where $P_w(\mathbf{red}) > .5$ is the same as the set of worlds where $P(\mathbf{red} \mid \mathbf{urn}) > .5$, relative to the actual value of **urn** at w. Relative to this model, the intension of *The ball is probably red*—$\{w \mid P_w(\mathbf{red} > .5\}$—is the same as the intension of *The Urn selected is A*. So, of course, conditioning on either has the same update effect.

We can now return to the puzzle around how to update with (3)/(4), *Team A is likely to win*. If we know nothing about these teams beyond general domain knowledge, our information might be represented by the Bayes net in Figure 2 above. On the present proposal the intension of *Team A is likely to win*, relative to this Bayes net, is $\{w \mid P_w(\mathbf{winner} = A) > .5\}$. The parents of **winner** in this model are $perf_A$ and $perf_B$. So, by the definition of P_w offered above, this is the set of worlds w where the probability of A winning is greater than .5, given the actual values of $perf_A$ and $perf_B$ at w. This result is problematic, since it implies that $P_w(\mathbf{winner})$ should be 1 or 0 at every world. This is because the value of **winner** is a deterministic function of its parents—**winner** $= A$ if $perf_A > perf_B$, otherwise $= B$. My suggestion is that we should look not necessarily to a variable's immediate parents, but to its closest non-deterministic parents.

> Final proposal: $P_w(V = v)$ is equal to the conditional probability of A given the actual values at w of V's closest non-deterministic ancestors in B.

On this proposal, the truth-value of *Team A is likely to win* at any w, relative to this model, is determined by the values of μ_A, σ_A, μ_B, and σ_B at w. These variables represent the facts at w about the teams' skill and consistency, respectively. Given the assumptions that we made in setting up the Figure 2 model, *Team A is likely to win* ends up having a sensible update effect: it is equivalent to $\mu_A > \mu_B$, i.e., the proposition that Team A is more skilled than Team B. Figure 5 shows graphically the effect of updating an uninformed prior distribution with this information.

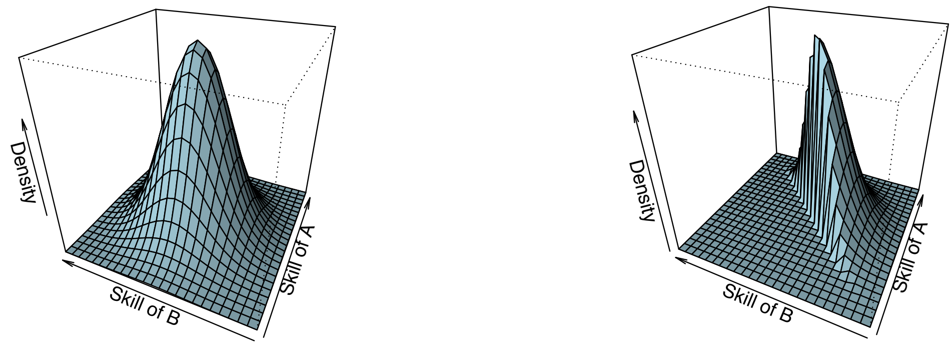


Figure 5: Left: 2-dimensional Gaussian representing an uninformed prior on the teams' skills. Right: posterior after learning that Team A is likely to win.

This proposal is tentative and subject to important caveats: I will mention only two for reasons of space. First, the final proposal might need to be circumscribed so that it applies only when neither V nor any of its descendants has been observed. After all, if we have observed that A won, then $P($**winner** $= A)$ should be 1. However, note that this observation could not have been made in the scenario under consideration, since the key example was a prediction about a future event which could not in principle have been observed. In general, a complete theory along these lines will need to be more explicit about the role of time—about the way that causal processes unfold in time, about the way that probability estimates change over time in response to new observations, and any interactions between the two. While I do not feel that I have a complete grasp of the complex issues at stake here, it does seem worth noting that it is often unproblematic to assert an explicitly tensed statement of the form ϕ *was likely and* $\neg\phi$—e.g., *Team A was likely to win, but they didn't win.* This is striking, since present-tense statements of the form ϕ *is likely, but* $\neg\phi$ are generally infelicitous (Yalcin, 2007). The difference in the past-tense case seems to be that there are two different times involved: an earlier time at which the "worldly" probability of A winning was greater than .5, e.g., because A is a more skilled team; and a later time at which the speaker learned that A didn't win, despite their advantageous position. So, we should not redefine P_w such that $P_w(V = v)$ is always 0 or 1 whenever the value of variable V has been observed. Instead, we need to explicitly incorporate an element of time: $P_{(w,t)}(V = v)$ is 0 or 1 if V has been observed at or before t. However, I will not attempt it in this space to working out the temporal aspect in detail or its connections to Yalcin's observations about "epistemic contradictions".

A second caveat is that my proposal may apply only to certain kinds of probability statements. Theoretical and corpus studies have found evidence that what we have called "probabilistic" language is ambiguous between multiple kinds of probability that are ontologically and psychologically distinct (Kahneman and Tversky, 1982). The probability statements that I have analyzed in the second part of this paper seem to relate to a kind of "objective", "worldly", or "stochastic" probability, relating to the unfolding of an indeterministic causal process. In contrast, most theorists in the semantic literature on probability expressions have explicitly or implicitly assumed that the subject matter of probabilistic language is *subjective* uncertainty. These positions are not in competition: Ülkümen et al. (2015), in particular, have shown that English has the semantic resources to talk about both kinds of uncertainty. They show that the two kinds of uncertainty are even distinguished lexically to some extent, with items such as *likely* and *chance* favoring stochastic interpretations while *confident*, *certain*, etc. favor subjective interpretations. Lassiter (2017b) gives further corpus evidence for this conclusion, and argues for several more semantically distinct interpretations of probability expressions.

The proposal floated here is most clearly suited to language describing stochastic causal processes, including (but not limited to) many future-oriented uses of probability expressions. I am exploring ways to extend it to clearly subjective uses like the corpus example in (6) (Lassiter, 2017b), where there is a clear fact of the matter and information is simply lacking.

(6) [T]he residential-scale reservoirs ... were likely used around 900 B.C. It's likely that the systems were lined with a thick, clay "plaster" ...

I am unsure whether this effort will be successful. I believe that it would be also be consistent with the proposal given here to model the update effect of such examples from a different perspective, such as Madsen's (2015) framework of "multi-agent statistics".

7. Conclusion

Imprecise credence models, which represent uncertainty using sets of probability measures, have a good deal of philosophical and linguistic motivation. However, their scope is more limited than has generally been recognized. Where the formal representation of uncertainty is concerned, imprecise credences may well be useful as a treatment of group belief. However, they encounter severe problems as a representation of individual-level uncertainty, particularly involving learning from a starting point of ignorance. I argued that these problems can be resolved by adopting an explicitly hierarchical perspective on belief states, and that hierarchical models already account for the confidence/ignorance distinction that has been used to motivate imprecise models as a representation of individual-level uncertainty. While I formalized this perspective using causal Bayes nets, there are also other, richer hierarchical models based on probabilistic programming techniques, which will probably be needed in a fuller treatment.

A second puzzle involving the update effects of expressions like ϕ *is likely* and *probably* ϕ

seemed initially to point to a formally parallel model, where information states are sets of measures (or something strictly richer). This model may well be sufficient as a common-ground model. However, it fails as a solution to the psychological problem of how agents can/should incorporate such statements into their information states, because it predicts that they should have trivial update effects when probabilities are known. I argued that we may be able to resolve this issue as well by attending to the hierarchical structure of belief states, interpreting at least some probabilistic statements as factual claims about stochastic causal processes rather than expressions of subjective uncertainty. While there is much more to be done to shore up this suggestion, I hope at least to have demonstrated that direct engagement between semantics and pragmatics, formal epistemology, computational cognitive science, and Bayesian statistics has significant potential to generate new concepts and useful theoretical models.

References

Bishop, C. M. (2013). Model-based machine learning. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 371*(1984).

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. MIT Press.

de Finetti, B. (1977). Probabilities of probabilities: a real problem or a misunderstanding? In A. Aykac and C. Brumet (Eds.), *New Developments in the Applications of Bayesian Methods*. North-Holland.

Elga, A. (2010). Subjective probabilities should be sharp. *Philosophers' Imprint 10*(5).

Gerstenberg, T. and N. D. Goodman (2012). Ping pong in church: Productive use of concepts in human probabilistic inference. In *Proceedings of the 34th annual conference of the cognitive science society*, pp. 1590–1595.

Glymour, C. N. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT press.

Goodman, N., V. Mansinghka, D. Roy, K. Bonawitz, and J. Tenenbaum (2008). Church: a language for generative models. In *Uncertainty in Artificial Intelligence*, Volume 22, pp. 23.

Goodman, N. D. and D. Lassiter (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin and C. Fox (Eds.), *Handbook of Contemporary Semantic Theory* (2 ed.). Wiley-Blackwell.

Goodman, N. D. and J. B. Tenenbaum. Probabilistic models of cognition. Retrieved January 6, 2017 from http://probmods.org.

Griffiths, T., J. B. Tenenbaum, and C. Kemp (2012). Bayesian inference. In R. G. Morrison (Ed.), *The Oxford handbook of thinking and reasoning*, pp. 22–35. Oxford University Press.

Griffiths, T. L., C. Kemp, and J. B. Tenenbaum (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Psychology*, pp. 59–100. Cambridge University Press.

Griffiths, T. L. and J. B. Tenenbaum (2006). Optimal predictions in everyday cognition. *Psychological science 17*(9), 767–773.

Halpern, J. Y. (2003). *Reasoning about Uncertainty*. MIT Press.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffrey, R. (1983). Bayesianism with a human face. In J. Earman (Ed.), *Testing Scientific Theories*, pp. 133–156. University of Minnesota Press.

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

Joyce, J. M. (2010). A defense of imprecise credences in inference and decision making1. *Philosophical perspectives 24*(1), 281–323.

Kahneman, D., P. Slovic, and A. Tversky (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kahneman, D. and A. Tversky (1982). Variants of uncertainty. *Cognition 11*(2), 143–157.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer.

Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. In N. Li and D. Lutz (Eds.), *Semantics & Linguistic Theory (SALT) 20*, pp. 197–215. CLC Publications.

Lassiter, D. (2011). Measurement and Modality: The Scalar Basis of Modal Semantics. Ph.D. thesis, New York University.

Lassiter, D. (2017a). Graded Modality: Qualitative and Quantitative Perspectives. OUP.

Lassiter, D. (2017b). Talking about higher-order uncertainty. Ms., Stanford University.

Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy 71*(13), 391–418.

Madsen, M. (2015). On the consistency of approximate multi-agent probability theory. *Künstliche Intelligenz 29*(3), 263–270.

Moss, S. (2015). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics 8*, 1–81.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pedersen, A. P. and G. Wheeler (2014). Demystifying dilation. *Erkenntnis 79*(6), 1305–1342.

Perfors, A. (2012). Bayesian models of cognition: What's built in after all? *Philosophy Compass 7*(2), 127–138.

Rinard, S. (2013). Against radical credal imprecision. *Thought: A Journal of Philosophy 2*(2), 157–165.

Rothschild, D. (2012). Expressing credences. In *Proceedings of the Aristotelian Society*, Volume 112, pp. 99–114. Wiley Online Library.

Seidenfeld, T. and L. Wasserman (1993). Dilation for sets of probabilities. *The Annals of Statistics*, 1139–1154.

Sloman, S. A. (2005). *Causal Models: How We Think About the World and its Alternatives*. OUP.

Stalnaker, R. (1978). Assertion. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics*. Academic Press.

Swanson, E. (2006). Interactions With Context. Ph.D. thesis, MIT.

Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman (2011). How to grow a mind: Statistics, structure, and abstraction. *Science 331*(6022), 1279–1285.

Trommershäuser, J., L. T. Maloney, and M. S. Landy (2008). Decision making, movement planning and statistical decision theory. *Trends in cognitive sciences 12*(8), 291–297.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*(4754), 1124–1131.

Ülkümen, G., C. R. Fox, and B. F. Malle (2015). Two dimensions of subjective uncertainty: Clues from natural language. To appear in *Journal of Experimental Psychology: General*.

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford University Press.

van Fraassen, B. C. (1990). Figures in a probability landscape. In J. Dunn and A. Gupta (Eds.), *Truth or consequences*, pp. 345–356. Springer.

Veltman, F. (1996). Defaults in update semantics. *Journal of Philosophical Logic 25*(3), 221–261.

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

White, R. (2010). Evidential symmetry and mushy credence. pp. 161–186.

Williamson, J. (2009). *In defence of objective Bayesianism*. Oxford University Press.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Yalcin, S. (2007). Epistemic modals. *Mind 116*(464), 983–1026.

Yalcin, S. (2010). Probability operators. *Philosophy Compass 5*(11), 916–937.

Yalcin, S. (2012). Context probabilism. In M. Aloni, V. Kimmelman, F. Roelofsen, G. W. Sassoon, K. Schulz, and M. Westera (Eds.), *Logic, Language and Meaning*, pp. 12–21. Springer.