

Chapter 2

Noisy Uncertainty

Abstract

Our subjective probabilities guide our behavior. So why can't we figure out what they are—removing our higher-order uncertainty—just by acting? Because *cognitive noise* makes the link between our subjective probabilities and our behavior stochastic. I summarize the empirical evidence for this, describe a popular model of it—the sampling hypothesis—and argue that reasonable agents will exhibit cognitive noise when and only when they have higher-order uncertainty. I report on an experiment supporting this prediction, and argue that Bayesian cognitive scientists have been using the wrong (implicitly, clear) models.

2.1 The Challenge

The main premise of this book is that reasonable people often don't know what their own opinions are, and that we should model this uncertainty the exact same way we model their uncertainty about anything else. To many that'll be as controversial as saying that reasonable people often don't know what their blood pressure is, and that we should model this uncertainty the exact same way we model uncertainty about anything else.

But—for some good reasons, and some bad—many theorists are skeptical of higher-order probabilities. Within philosophy, this skepticism often comes from an internalist impulse: if we don't know what we think, how can it guide our behavior? Outside of it, it often comes from a behaviorist impulse: if your actions are coherent, we can impute a probability function that represents them; how could the higher-order uncertainty matter?

The skepticism is bolstered by famous arguments claiming to show that higher-order probability is either incoherent or trivial (e.g. Savage 1972, §4.2; de Finetti 1974, §4.9). And while it's easy to write down abstract models showing that there's no *mathematical* incoherence or triviality, it's much harder to say how to interpret those models. What's needed is an interpretation that explains (1) what it means for those models to represent your true (higher-order uncertain) probabilities, (2) how your (uncertain) probabilities could nevertheless influence your actions, and (3) what prevents

you from using such influences to *remove* your higher-order uncertainty by observing how you act.

In this chapter I'll offer such an interpretation. The key fact? We suffer from **cognitive noise**—randomness or stochasticity in the link between our mental states and actions. Higher-order uncertainty turns out to be the natural state of a noisy probabilistic system. I'll illustrate these arguments with the *sampling model* of subjective probabilities that's popular throughout cognitive science. I'll close by arguing that *any* reasonable agent whose subjective probabilities suffer from cognitive noise must have higher-order uncertainty, and leverage this into a critique of many Bayesian models used in cognitive science.

My goal here is not to defend the rationality of higher-order uncertainty—that'll come in Chapters 5 and 6. But I do want to defend its mathematical and conceptual coherence. For those impatient for the mathematics, the following (optional) subsection write down a simple model of higher-order uncertainty of the sort I'll use throughout this book. For those impatient get to the conceptual questions, skip to §2.1.2.

2.1.1 Mathematical Coherence[†]

Using the right tools, it's easy to show the mathematical coherence of higher-order uncertainty. We'll do this slower and in more detail in Chapter 3 and Part II; here I'll be brief.¹

A *probability frame* (W, \mathbf{a}, P) consists of a (say, finite) set of worlds W , an actual world $\mathbf{a} \in W$, and a function from worlds w to probability distributions P_w over (all subsets of) W . P is a random variable thought of as a *description* of a probability function—like 'your subjective probability function, whatever it might be'. P_w , in contrast, is a constant, (rigidly) designating P 's realization at world w . P has higher-order uncertainty at a world w iff P_w leaves open worlds v and u where P has different values: $P_w(v) > 0$ and $P_w(u) > 0$ but $P_v \neq P_u$. Since W is fine-grained enough that worlds specify the values of P , we can thereby use P to define propositions about your credences as sets of worlds (events) in the frame—for example, $\langle P(q) = 0.55 \rangle := \{w \in W : P_w(q) = 0.55\}$ is the set of worlds at which P assigns 0.55 probability to q . Such definitions let us to 'unravel' higher-order probabilities to be simply probabilities of events—for instance, $P_w(\langle P(q) = 0.55 \rangle) = 0.6$ is true iff P_w assigns 60%-probability to the event $\{x \in W : P_x(q) = 0.55\}$.

Here's an example. Suppose you are unsure whether you are 0.45 or 0.55-confident that I own a dozen spoons. Then you need to leave open two classes of worlds—one where $\langle P(d) = 0.45 \rangle$ and one where $\langle P(d) = 0.55 \rangle$. Let's suppose—unbeknownst to you—that in fact you assign (subjective) probability 0.55 to d . Then the actual world \mathbf{a} is in $\langle P(d) = 0.55 \rangle$, and $P_{\mathbf{a}}(d) = 0.55$. But since $P_{\mathbf{a}}$ leaves open both classes of worlds, $P_{\mathbf{a}}(P(d) = 0.45) > 0$ and $P_{\mathbf{a}}(P(d) = 0.55) > 0$.²

Let's write down a toy model that makes these claims true. Cross the questions of whether I own a dozen spoons (d) or not (\bar{d}) with whether your credence that I do is *low* (l) or *high* (h). That yields 4 possible worlds; let's list them in the order $(d_h, \bar{d}_h, d_l, \bar{d}_l)$. To specify a probability frame, we need to write down a function from these 4 worlds to probability distributions *over* those 4 worlds. A probability distribution over a finite set of worlds can be written as a non-negative

¹The foundational work is in Kripke 1963; Hintikka 1962, and Harsanyi 1967. For uses of such models, see e.g. Gärdenfors 1975; Gaifman 1988; Samet 2000; Williamson 2000, 2008, 2014; Schervish et al. 2004; Lasonen-Aarnio 2015; Salow 2018; Das 2022; Dorst 2020a; Dorst et al. 2021; Dorst 2023b.

²When events like $\langle P(d) = 0.45 \rangle$ are embedded, I'll often omit the angle brackets for readability.

vector whose entries sum to 1; for example, $\pi_h = (0.45, 0.15, 0.1, 0.3)$ assigns 0.45 to d_h , 0.15 to \bar{d}_h , etc. Probabilities of events are obtained by summing across the worlds in which they're true, so $\pi_h(d) = \pi_h(\{d_h, d_l\}) = 0.45 + 0.1 = 0.55$.

Using these conventions, Figure 2.1 shows a full frame. The first row and column are labels. The row that starts with d_h specifies your credence distribution P_{d_h} at d_h , saying that you have π_h —assigning 0.45 to d_h , 0.15 to \bar{d}_h , etc. Since this distribution assigns 0.55 to d , and you have the same distribution at \bar{d}_h (the row below), the claim $\langle P(d) = 0.55 \rangle$ is equivalent to $\{d_h, \bar{d}_h\}$ in this frame. Similarly, at d_l you have distribution $\pi_l = (0.3, 0.1, 0.15, 0.45)$, assigning $0.3 + 0.15 = 0.45$ to d . Since you have this same distribution at \bar{d}_l , the claim $\langle P(d) = 0.45 \rangle$ is equivalent to $\{d_l, \bar{d}_l\}$.

$$P = \left(\begin{array}{c|cccc} & d_h & \bar{d}_h & d_l & \bar{d}_l \\ \hline \mathbf{d}_h & 0.45 & 0.15 & 0.1 & 0.3 \\ \bar{d}_h & 0.45 & 0.15 & 0.1 & 0.3 \\ d_l & 0.3 & 0.1 & 0.15 & 0.45 \\ \bar{d}_l & 0.3 & 0.1 & 0.15 & 0.45 \end{array} \right)$$

Figure 2.1: A model of ambiguous (higher-order uncertain) opinions about whether I own a dozen spoons (d).

I've specified the actual world $\mathbf{a} = d_h$ by bolding it. Thus, in fact, your actual credence in d is $P_{\mathbf{a}}(d) = 0.55$. But you are, in fact, unsure about this: you assign only $0.45 + 0.15 = 0.6$ -probability to $\langle P(d) = 0.55 \rangle = \{d_h, \bar{d}_h\}$, and so assign $0.1 + 0.3 = 0.4$ -probability to $\langle P(d) = 0.45 \rangle = \{d_l, \bar{d}_l\}$. Despite being 0.55-confident of d , you are unsure whether you are 0.55 or 0.45-confident of it. You have genuine higher-order uncertainty.

This model embeds many mathematical and conceptual subtleties. The point for now is simply that it is a mathematically-coherent model of higher-order uncertainty: at every world, you have a probabilistically-coherent probability function that assigns exact values to every relevant proposition, while also having (probabilistic) uncertainty about what your probabilities are.

2.1.2 Conceptual Coherence

So higher-order uncertainty is mathematically coherent. Is it also *conceptually* coherent? In my experience, writing down models like this rarely moves the needle—people remain skeptical of how such a model could correctly encode your opinions. So, how could it? Start with the obvious.

First point: *I* can have be unsure what your opinions are. To model my uncertainty about your opinions, we must treat your credences P as a *random variable* or a *description* which picks our different probability functions in different possibilities. First upshot: there are facts about what your opinions are, which can be modeled with a variable that varies across possibilities.

Second point: you have *opinions* about your own opinions. Do I own at least one spoon? Of course. Are you confident that I do? *Also* of course. You assign probability (near) 1 to the claim that *you're over 90%-confident that own at least one spoon*: $P(P(\geq 1 \text{ spoon}) > 0.9) \approx 1$. So facts about your opinions are in the domain of the probability function representing your beliefs—you're have opinions about your own opinions. Second upshot: the only way to avoid higher-order uncertainty is to be higher-order *certain*.

Third point: cognitive noise—imperfect reliability in eliciting your opinions—makes it hard to

figure out exactly what your opinions are. If I ask you how confident you are that I own a dozen spoons, and you say ‘55%’, I won’t bet the farm that you’re 55%-confident. After all, I suspect you’re not perfectly reliable at reporting your opinions, just as you’re not perfectly reliable at reporting your moods. The same goes for you. Suppose you don’t know whether you’re more or less than 50%-confident in d . You might try to find out by asking yourself, ‘Would I rather bet on d or $\neg d$?’. But you might well be unsure which you prefer. You could force yourself to choose, but then you’ll wonder whether that choice reflects your true preferences. Third upshot: cognitive noise makes it hard to be sure what your (own) opinions are.

Cognitive noise is randomness or stochasticity in the causal path between a mental state (like a credence or preference) and how it manifests (like a decision to bet on d , or to announce ‘55%’ as your probability). It’s familiar in other contexts. Last night I felt a bit crummy. Why? That was the question. Maybe I was stressed about work; maybe I was just dehydrated. These hypotheses called for different actions: if stressed about work, I should write a plan for the week; if dehydrated, I should chug a glass of water. What to do? I tried asking myself: ‘Am I stressed?’ But it didn’t work: introspection yielded an unhelpful, ‘...maybe?’. When I forced myself to guess, I said ‘Yes’. But I wasn’t sure that was right—the link between whether I’m stressed and whether I *say* I’m stressed is tenuous. I decided to chug the glass of water and waited 20 minutes. It worked.

No surprises here. We know that many of our mental states are both hard to introspect and unreliable in their manifestations. What we need is to understand how this could be true of our (probabilistic) *beliefs* or *credences*.

Credences (subjective probabilities) often live under the shadow of behaviorism: ‘Your credences are revealed by your actions’, it’s said. Obviously there’s something to this. Credences are functional states—if some module in your brain spits out probabilities that don’t affect your actions, they aren’t your credences. But equally obviously, this ‘revealed confidence’ picture (like ‘revealed preferences’) is too crude—sometimes we take risks that weren’t worth it, even by our own lights.³

What we need is a clear understanding of how your credences could be linked to your actions, without being *perfectly* so-linked. The answer—as psychologists have been telling us for some time (Thurstone 1927)—is that the link can itself be probabilistic. Being 55%-confident of d *tends* to lead you to prefer to bet on d rather than on $\neg d$, *tends* to lead you to write down numbers near ‘\$0.55’ for your fair betting price, and so on. But that tendency is itself chancy and subject to noise. Let’s see how this could work.

2.2 The Sampling Model

Cognitive science contains many models on which subjective probabilities are noisily linked to action (see Weisberg 2020). I’ll focus a popular one: the **sampling model** (Icard 2016)—or, as I prefer to call it: *urns in the head*. I’ll use it throughout the book to ground ideas, so it’s worth

³Maybe your credences are revealed by your actions *in ideal conditions*? But no, that’s a distraction. Our lives are spent in non-ideal conditions—wherein our opinions are subject to noise, self-doubt, and so on—and we want to understand how it’s reasonable to think and act *given those conditions*. To idealize them away would be to change the subject. Ideal-conditions analyses of credences fail in the way that most counterfactual analyses do—by failing to track how the antecedent of the counterfactual can affect the state they’re analyzing (see Williamson 2000, Ch. 10).

understanding. But it's not essential to any of the main arguments. Although I think it has a lot going for it, nothing hinges on it being an accurate model of real people's subjective probabilities.

Here's how it works. There's both theoretical and empirical reason to think that the mind—and perhaps any domain-general probabilistic reasoner—often tracks probabilities as follows. We have a *generative model* of a given subject-matter \mathcal{Q} that can simulate what might be true about it, i.e. generate possible answers to questions about \mathcal{Q} . This model will work by implicitly encoding general dynamics—like intuitive laws of physics, or how people's beliefs and desires influence their behavior. When those dynamics are combined with information about particular scenarios, they can then be used to simulate what might happen—if you throw the rock *this* hard, or you try to talk with your uncle about *that*, what happens next?

Fixing a given scenario, subject-matter, and instant in time⁴, your generative model has certain *sampling propensities*—dispositions to generate hypothetical outcomes—and you can *sample* from ('simulate') it to probe these propensities and let them influence your behavior. Think of it like drawing marbles from a complicated, opaque urn inside your head, with a different type of marble for each possible answer to \mathcal{Q} . Your credences P about question \mathcal{Q} are the sampling propensities of your model—the limiting results, were you to sample infinitely (without changing your model). The (finite) samples you actually draw, ϵ , are the **elicitations** of these propensities; these elicitation are what directly determine your behavior in a given moment.

For example, suppose you construct a model about whether I own a dozen spoons. This'll be a largely subconscious process that works like a computer program that uses some relevant information—about my demographics, culture, possible spoon-obsessions, etc.—to generate 'samples', i.e. possible answers to the question \mathcal{Q} . It'll be stochastic (in its dynamics or initial conditions), so will generate different samples on different runs. In our case, the answers to \mathcal{Q} will take a stand on whether I own a dozen spoons (as well as, perhaps, other questions—like exactly how many spoons I own). Your credence that I do, $P(d)$, equals the objective chance that your model has of generating a ***d*-sample** (a sample in which *d* is true).

If you draw one sample, that'll be very noisy, and may be unrepresentative of what's likely, given your model. So instead you could run the model many times, generating many samples. You can then use those samples to guide your behavior—for example, using facts like which outcome is most common in your samples, or what proportion of them were *d*-samples, to guide your behavior. If real people do this, obviously it's subconscious—we may be dimly aware that we are imagining ('simulating') various possibilities, but we aren't consciously counting them. As with the other everyday feats of human cognition—like walking, talking, and seeing—we are not conscious of the impressive computational work underlying our imagination.

Here's a toy example. Suppose the subject-matter \mathcal{Q} consists of two questions: (1) whether I like cereal (*c*) or not (\bar{c}), and (2) whether I own a dozen spoons (*d*) or not (\bar{d}). There are four possible outcomes: cd , $c\bar{d}$, $\bar{c}d$, $\bar{c}\bar{d}$. Suppose that the sampling propensities of your model are the same as those of drawing with replacement from an urn with marbles labeled ' cd ', ' $c\bar{d}$ ', etc., with following numbers of each:

⁴Your sampling propensities will be constantly changing, both from new information and (as we'll see) by updating on the samples generated by your model itself. But set that aside for now.

2.2. THE SAMPLING MODEL

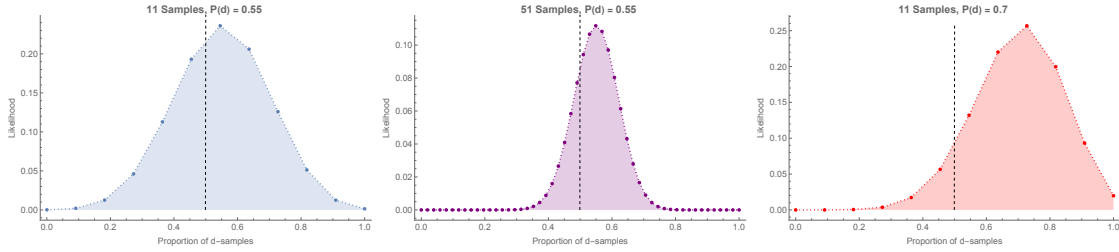


Figure 2.2: *Left:* Likelihood of drawing a given proportion of d -samples out of 10 total, if $P(d) = 0.6$. *Middle:* Likelihoods for 50 samples. *Right:* likelihoods for 10 samples if $P(d) = 0.7$.

| cd | $c\bar{d}$ | $\bar{c}d$ | $\bar{c}\bar{d}$ |
|------|------------|------------|------------------|
| 35 | 15 | 20 | 30 |

Then your credence that I own a dozen spoons is the probability that you’ll draw a d -sample: $P(d) = \frac{35+20}{35+15+20+30} = \frac{55}{100} = 55\%$. Similarly for conditional credences: your conditional credence that I own a dozen spoons, given that I like cereal, is the proportion of marbles that say *I like cereal* (cd or $c\bar{d}$) that are ones where I own a dozen spoons: $P(d|c) = \frac{35}{35+15} = 70\%$.⁵

Now suppose I make you guess whether I own a dozen spoons. On the sampling model, this involves eliciting your credence by sampling, and then acting on those samples. Suppose you draw 11 samples, record the proportion that are d -samples, and guess d iff more than half of them are d -samples. This is statistically identical to flipping a coin 11 times that’s 55%-likely to land heads; the likelihoods are a Binomial(11,0.55) distribution (Figure 2.2, left).

Suppose—as is around 24% likely—6 of your samples are d possibilities, and 5 are $\neg d$ ones. This leads to an elicited distribution ϵ such that $\epsilon(d) = \frac{6}{11} \approx 55\%$ -confident of d and $\epsilon(\neg d) = \frac{5}{11} \approx 45\%$ -confident of $\neg d$. This matches your true underlying credences, and since ϵ thinks d is more likely than $\neg d$, it leads you to guess that d —your action matches your underlying credal state.

But it’s perfectly possible that *fewer* than 6 of your samples are d -samples—indeed, it’s around 37%-likely. (Sum the likelihoods below dashed line.) If it does, then despite the fact that you’re actually 55%-confident that I used a spoon ($P_a(d) = 0.55$), your elicited distribution is more confident I *didn’t* use a spoon ($\epsilon(d) < 0.5$, so $\epsilon(\neg d) > 0.5$). Thus it leads you to guess that I *didn’t* use a spoon—noise has distorted the channel from your credence to your action, leading you to act in ways that don’t match your own underlying mental states. How to make such distortions less likely? Two obvious ways: you could draw *more samples* (Figure 2.2, middle), or you could become *more confident* of d (Figure 2.2, right).

Let’s generalize this. The sampling model assumes some sampling propensities P over an (ordered) set of possibilities (w_1, \dots, w_n) —in our case, $(cd, c\bar{d}, \bar{c}d, \bar{c}\bar{d})$. When you elicit them, you draw k samples from P , recording what proportion of the samples were each type. Think of this like drawing (with replacement) from an urn with n different types of marbles in it, with the proportion of w_i -marbles in the urn matching the sampling propensity $P(w_i)$. Letting k_i be the number of samples in which w_i was drawn (so $k = k_1 + \dots + k_n$), your **elicited probability**

⁵Thinking in terms of urns, sampling is done *with replacement* to ensure that it doesn’t change the generative model, so samples are independent given the model. If you like, think of re-running a computer program. We’ll see later how sampling might change your opinions if you update on your own elicitation.

distribution ϵ over (w_1, \dots, w_n) is $\epsilon = (\frac{k_1}{k}, \frac{k_2}{k}, \dots, \frac{k_n}{k})$. For example, suppose you draw 11 samples and 3 are cd -samples, 2 are $c\bar{d}$ -samples, 2 are $\bar{c}d$ -samples, and 4 are $\bar{c}\bar{d}$ -samples. Then your elicited distribution is $\epsilon = (\frac{3}{11}, \frac{2}{11}, \frac{2}{11}, \frac{4}{11}) \approx (0.27, 0.18, 0.18, 0.36)$, differing slightly from your true underlying probabilities of $P = (0.35, 0.15, 0.2, 0.3)$.⁶ On the sampling model, the elicitation ϵ is a transient state that guides your action in a given moment—for example, you might maximize expected utility according to it.⁷ You draw samples *so that* your underlying beliefs can (noisily) influence your actions. In this case, if you use ϵ to guess whether I used a spoon, you’ll guess that I *didn’t* (since $\epsilon(d) = \frac{5}{11}$), despite the fact that your true credences think it’s more likely that I did.

In short, the noise between your credences (P) and your elicitation of them (ϵ) is what allows you to act in ways that don’t align with your beliefs. This makes an important separation. Classically, there are two roles for credences: (1) directly encoding your beliefs and your responses to evidence (via conditioning), and (2) directly interfacing with action (via maximizing expected utility). The sampling model separates these two roles: it is your sampling propensities (P) that directly encode your beliefs and respond to evidence, while it is your elicitations (ϵ) that directly guide your behavior. The two are linked by a noisy channel. I’m making a terminological choice when I used ‘subjective probability’ for (1) rather than (2); but what matters is just that to understand cognitive noise between belief and action, we *need* to separate the two

Or so I claim. I’ve assumed that cognitive noise involves a (relatively) stable mental state that’s distorted by noise when it gets elicited for action—call this the ‘elicitation theory’. But there’s an alternative—call it the ‘fluctuation theory’—on which there’s no stable mental state. Rather, there is just your constantly-fluctuating elicitation: your *true credence* that I own a dozen spoons is 60% one second, 43% the next, 71% the next, and so on (cf. Easwaran 2024).

There’s less daylight between elicitation- and fluctuation-theories than it might seem. First, elicitation-theories agree that your underlying sampling propensities *do* constantly change: as you bring new evidence to bear (including from the results of sampling) your generative model will shift. And fluctuation-theories also predict that higher-order uncertainty persists after acting. Suppose you guess that I *do* own a dozen spoons. Two seconds later, are you certain that you’re more than 50%-confident of it? No. On the fluctuation theory, you are unsure whether your quickly-oscillating opinion is *still* above 50%—for all you know, it’s dropped since you guessed.

Still, I prefer elicitation-theories. They are simpler to work with, mathematically and computationally. They are more rational—fluctuation theories require your credences to be constantly shifting wildly on the basis of little or no evidence. And fluctuation theories seem to be confusing your beliefs for your *occurrent* beliefs. We have many beliefs that aren’t conscious, but explain our dispositions.⁸ Even when I haven’t thought about my mom for days, I still believe (know!) that her birthday is March 10th. That belief explains why noticing that it’s March 6th spikes my heart rate, reminding me to buy a gift. And even if—due to a momentary lapse—I consider buying her the

⁶In the finite case, you’re sampling from a ‘multinomial distribution’—a generalization of the binomial. We can recover those binomial probabilities by lumping together outcomes that agree on whether d is true—if we’re just tracking whether or not it’s an d -sample, drawing from the four-place probability vector $(0.35, 0.15, 0.2, 0.3)$ over $(cd, c\bar{d}, \bar{c}d, \bar{c}\bar{d})$ is equivalent to drawing from the two-place vector $(0.55, 0.45)$ over (d, \bar{d}) .

⁷How long does ϵ last, before you have to re-draw samples to act on your credences? I don’t know. There’s some evidence for ‘arbitrary coherence’—that people’s initial estimates are arbitrary but that over the course of a short study they coherently follow through on them (Arieli et al. 2003).

⁸See Ryle 1949; Anderson 1983; Pylyshyn 1984; Stalnaker 1984; Fodor 1987; Dennett 1989.

box of caramels I'm holding, I still believe (know!) that she doesn't like caramels. Similar lessons apply to your fluctuating, occurrent assessment of how likely I am to own a dozen spoons.

So let's stick with the elicitation-based sampling model. Five points about it.

First: it isn't an *analysis* of (noisy) credences; it's a hypothesis about how they're realized.

Second: as the number of samples (k) grows, it's increasingly likely that the elicitation ϵ will closely approximate your true credences P .

Third: cognitively-realistic generative models will be over restricted domains, so your individual models might not 'hang together' in a probabilistically-coherent way. We *could* explicitly model this.⁹ But my focus is on higher-order uncertainty, which doesn't require logical non-omniscience or probabilistic incoherence (see footnote 14 below). So let's assume your cognitive models hang together in a coherent overall distribution, P .

Fourth: when using the sampling model to interpret your credences P , I'll assume that you know your true sampling distribution (credences) P are rational, but may (under ambiguity) be unsure what they are. This is a choice point that nods to a subjectivist-Bayesian approach to rationality that might make certain philosophers unhappy. After all, they'll say, you might perfectly-well know that you're 60%-confident that I own a dozen spoons, and still wonder if that's the *rational* credence to have. Don't we want higher-order uncertainty to model this sort of normative uncertainty?

I used to agree (Dorst 2020a), and I still think such purely-normative uncertainty is perfectly legitimate. But I've come to think that it makes higher-order uncertainty more complicated mathematically, and needlessly mysterious conceptually—especially to those less comfortable with appeals to unanalyzed normative concepts. If you know what your *actual* credences are, and are just uncertain what the *rational* ones are, then we have multiple probability functions *for you, now* floating around, and our model needs to explain how they are linked. And the imagined scenario is unrealistic anyways: in the real-world cases of ambiguity, we *don't* know what our actual credences are—you genuinely are unsure what you think about my spoon collection, and modeling you as knowing what you think but not knowing what you *should* think is distorting. Finally, my preferred interpretation can still account for the intuitions about the case. On that interpretation, you standard know exactly what your *elicitation* is—you just wrote down '55%', after all—and are unsure whether *that* was rational, since you're unsure whether it aligns with your true opinions.

Regardless, the main arguments of this book don't rely essentially on the sampling model; analogous points could be developed where P is interpreted as the *rational* credence function for you, and not also your actual credence function. The point of the sampling model is to provide a tractable-but-realistic model of how you could be uncertain what your (rational) opinions are, and they could still influence your action.

Fifth and final point: noisy-sampling is sometimes said to make bias inevitable (e.g. Woodford 2020). This may sound like my claim (§1.4) that *higher-order uncertainty* makes biases inevitable. It's not. Most noisy-sampling models in cognitive science are—mistakenly, in my view—noisy approximations of Standard-Bayesian models (see §2.5 below). As a result, they are noisy approximations of a set of opinions P that converge to the truth. The 'biases' that such noise induces are *statistical* biases—predictable deviations between your elicitations and your true subjective prob-

⁹See the literature on fragmentation and similar models: Lewis 1982; Stalnaker 1984; Cherniak 1990; Greco 2015; Yalcin 2018; Staffel 2020; Elga and Rayo 2021, 2022; Hoek 2021, 2022; Koralus 2023.

Not terribly happy with this.

abilities P . In contrast, when I say that higher-order uncertainty makes biases inevitable, I mean that it makes it so that *the rational credences themselves*, P , are biased in the sense that they can predictably fail to converge to the truth.¹⁰

2.2.1 The Empirical Evidence

What’s the evidence for the sampling model? Though broad, it’s also a bit diffuse (see Icard 2016; Sanborn and Chater 2016; Zhu et al. 2023 for summaries). Here are some highlights.

On the theoretical side: exact probabilistic inference is computationally intractable (Cooper 1990; Dagum and Luby 1993), and most ways of approximating it use ‘Monte Carlo’ methods that sample from the relevant probability distributions (MacKay 2003; Gelman et al. 2014; McElreath 2018). One of the first lessons of programming is that if you try to explicitly calculate probabilities, your computer will give up after a few dozen coin-tosses—there are just too many possibilities whose probabilities all need to be multiplied and summed. Yet it’s trivial to get accurate estimates of probabilities just by simulating them and counting the results.

Next, empirical evidence. First: in some domains—like visual, physical, and social reasoning—it’s widely accepted that people use generative models to form their judgments (Marr 1982; Spelke et al. 1992; Lake et al. 2017). For example, take a look at the block tower in Figure 2.3. Do you think it’s stable, or will it fall? How confident are you? In answering the first question, it’s widely accepted that you have a generative (physical) model of the blocks, encoding your implicit knowledge of physics—think of it like a video game engine—which you use to generate your verdict.¹¹ Since the parameters of this simulation are uncertain (‘How *exactly* are the blocks arranged?’), there’ll be randomness in the outcomes. As a result, it would be natural for your mind to use the outcomes of multiple simulations to modulate its confidence—and indeed there’s evidence that it does (e.g. Battaglia et al. 2013; Smith et al. 2024). This mechanism can also explain other findings, such as multi-stable visual scenes like the Necker Cube (Figure 2.3, middle)—different samples will make different physical interpretations more likely (Gershman et al. 2012).

Second: people often *probability-match*, choosing an option in proportion to how likely it is to be best. If they’re observing the outcomes of a coin that’s in fact 70%-biased toward heads, they’ll end up predicting heads 70% of the time and tails the other 30% (Vulkan 2000). This is suboptimal—it’s better to predict heads *every* time. Yet probability-matching is what the sampling model predicts if you draw only one sample: 70% of the time, the sample will be *heads*.

In binary-choice settings, there are other explanations of (approximate) probability-matching—for example, if people elicit their probabilities with noise and then maximize expected value. But

¹⁰For aficionados: the method of sampling I described—wherein we draw directly from P —is called ‘rejection sampling’. It’s a ‘statistically unbiased’ estimator of P : conditional on $P(q) = x$, the best estimate for $\epsilon(q)$ is x . The cost of being unbiased is that rejection sampling is inefficient: if e is unlikely enough, it may require *many* samples to get an accurate estimate of $P(q|e)$. Thus Bayesian statisticians and psychologists often use sampling methods (like ‘Markov Chain Monte Carlo’) that are more efficient, at the cost of being (in the short run) biased estimators of P : their elicitation $\epsilon(q|e)$ will, in expectation, under- or over-estimate the true value $P(q|e)$ (see e.g. Lieder et al. 2012, 2018; Icard 2016). I’m not denying that this is a source of bias. What I’m claiming is that higher-order uncertainty induces a *further*—I think, deeper—source of bias in the rational credences (P) themselves.

¹¹E.g. McCloskey 1983; Spelke 1990; Tenenbaum et al. 2011; Battaglia et al. 2013; Ullman et al. 2014; Ullman and Tenenbaum 2020.

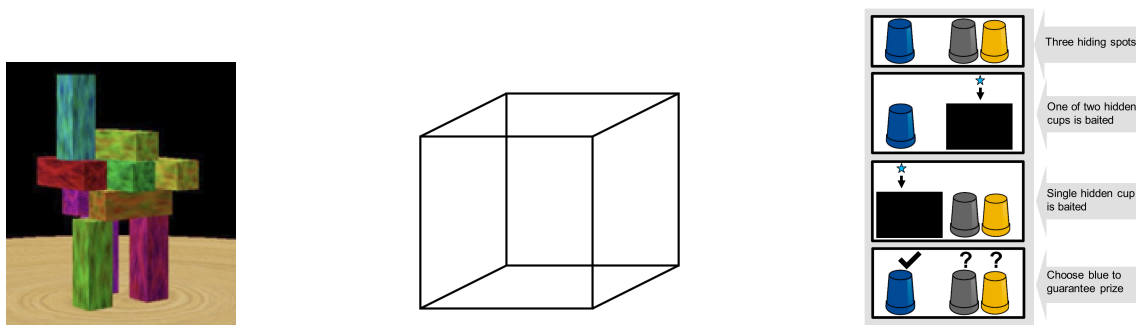


Figure 2.3: *Left:* Your intuitive judgments about the block-tower’s stability illustrates your implicit use of generative (physical) models. *Middle:* repeated sampling from uncertainty about the physical parameters can explain why your interpretation of the Necker cube bounces back and forth between two alternatives. *Right:* The 3-cups task, which provides evidence that children and chimps run (single) simulations.

these alternatives break down with more than two outcomes. A striking illustration comes from the ‘3-cups task’—see the right of Figure 2.3. A lone cup is on the left, and a pair of cups are on the right. A treat—a sticker for children; a snack for chimps—is put in one of the two right cups, behind an occluder so that subjects are unsure which one. Then another treat is put in the leftmost cup. Picking the left cup is guaranteed to give a treat, while picking one of the right two has a 50% shot. Subjects who always maximized expectation would always pick left, while subjects who randomized would pick each cup with $1/3$ probability. Preschoolers and chimps do neither: instead, they pick the left cup 50% of the time, and each of the right cups 25% of the time (Hanus and Call 2014; Mody and Carey 2016). Why? Suppose that they’re running a single simulation and acting on it. The simulation will always yield a treat in the left cup, and will yield a treat in each of the right two cups half the time. If they then take the best option *given* this simulation, they’ll randomly choose between the two locations that the simulation says contains a treat—which, across subjects, will lead to the observed 50-25-25 choice behavior (Leahy and Carey 2020; Leahy 2023, 2024, cf. Phillips and Kratzer 2024).

Third: people don’t *exactly* probability-match. When the stakes go up, they’re increasingly likely to choose the expectedly-best option (Zhu et al. 2023). The sampling model predicts this, assuming higher stakes lead people to draw more samples. In fact, such deviations from probability-matching are well-explained if sampling is costly, so that people only sample more when it’s worth it (Vul et al. 2014). More generally, people robustly exhibit a *speed-accuracy tradeoff* in their judgments: faster judgments tend to be less accurate (Garrett 1922; Johnson 1939). This is predicted by the simple fact that drawing more samples leads to greater accuracy but takes more time (Zhu et al. 2023).

Finally: when people deviate from probability-matching, their response rates often generate a ‘sigmoid’ (S-shaped) curve around the point of equal expected value (Woodford 2020). Figure 2.4 (left) shows this *for a single person* when repeatedly given the option to pay 5 cents for gambles that had a 50%-chance to pay x cents, as x varies, the probability of taking the bet follows an S-curve centered near the point of equal expected value (Mosteller and Noguee 1951). This is predicted if people maximize expected value with respect to their elicited samples (Figure 2.4, right), since more samples are likely to have an average value close to the expected value.

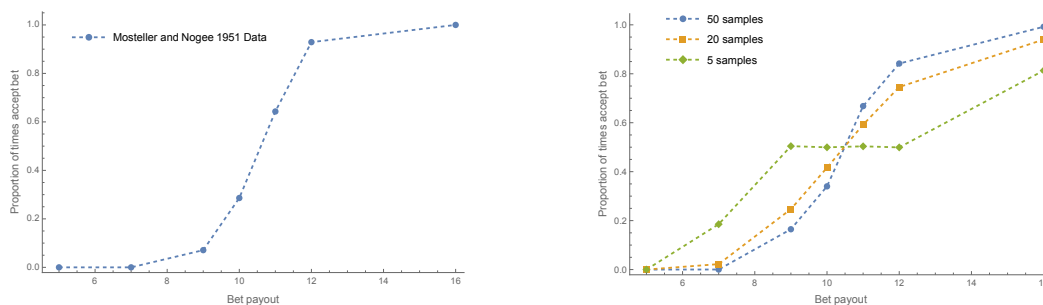


Figure 2.4: Noisy, S-shaped empirical response rates (left) are predicted by the sampling model (right).

Sampling also makes intuitive sense. Consider how we make social predictions. How would my brother react if I asked to borrow his car for a week? I’ve never done that; but I can imagine what he might do, say, or ask. At some level of description, my mind is clearly simulating what might happen based on what I know about Chris, our relationship, and so on, making explicit my implicit knowledge of these things.¹²

In short: sampling is a theoretically-motivated, empirically-supported, and intuitively-plausible hypothesis about how creatures like us might track and act on our uncertainties.

2.2.2 Sampling and Higher-Order Uncertainty

That’s sampling. On this model, it’s intuitive that you can have higher-order uncertainty. After all, P is a well-defined probability function that’s difficult to discern.¹³ Think of a computer program that generates a random number between 10–20, another between 5–10, and then outputs their product:

```

Generative Model 1:
x1 = RandomReal[{10,20}];
x2 = RandomReal[{5,10}];
Return[x1·x2]

```

This determines a precise likelihood distribution over outcomes. So... what is it? What’s its mean, or median? How likely is it to generate a number below 75? We don’t know.¹⁴ The easiest way to

¹²Philosophers have recently gotten interested in the way we use our imagination to come to know modal or counterfactual truths; what they have in mind is a very similar process (Williamson 2007, 2020; Kind and Kung 2016; Badura and Kind 2021; Myers 2021).

¹³Sampling propensities could well be inexact in the sense of not picking out a real-valued probability (Chapter 1, footnote 6). I’ll ignore this complication.

¹⁴Does higher-order uncertainty require logical non-omniscience? No. Of course, if you know the algorithm that generates your samples—as you do here—then being unsure of what distribution it generates does require logical non-omniscience. There are many ways to model this (cf. Lewis 1982; Garber 1983; Stalnaker 1984, 1991; Égré 2020; Elga and Rayo 2021; Hoek 2022). All will involve possibilities where the model generates a different distribution. All will be quite subtle. Happily, we can avoid these subtleties: it’ll work equally well to say that you *don’t know* exactly what algorithm generates your samples—it’s not as if you can peak inside and see what’s written on your neurons. Just as you can have (and manipulate) an urn without knowing what’s in it, likewise you can have (and manipulate) a generative model without knowing what algorithm it’s running. We can model P as logically omniscient but higher-order uncertain.

2.2. THE SAMPLING MODEL

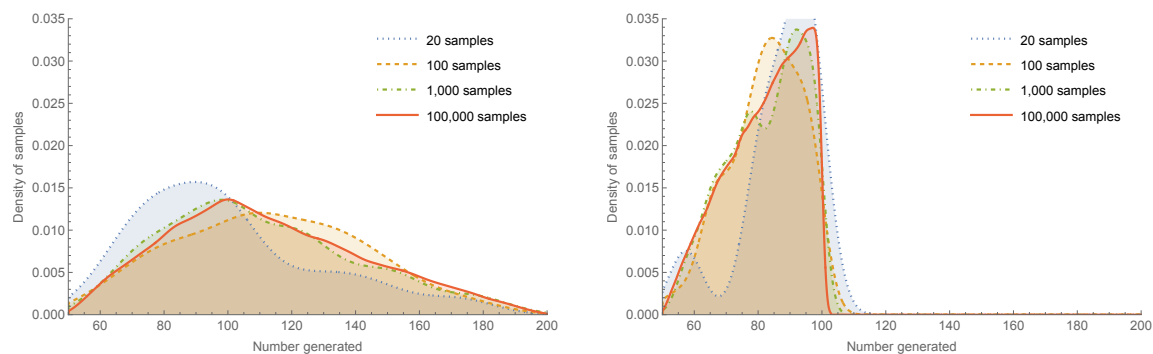


Figure 2.5: Samples from Generative Model 1 (left), and the result of the result of conditioning it on 0–100 (right).

find out is to *sample*: run the program and record its outputs. Figure 2.5 (left) shows the results of varying numbers of samples. Its mean is around 112, its median is around 109, and it’s roughly 11%-likely to generate a number below 75. Notice that with 20, 100, or even 1000 samples, we can easily mis-estimate the true distribution.

So we’re often unsure of a model’s sampling propensities. Likewise, we can *change* it in a definite way and be unsure how this affects its propensities. For example, we can condition it on the outcome being between 0–100, adding a `While`-loop that re-runs the algorithm if it’s greater than 100:

```
Generative Model 2:  
x1 = RandomReal[{10,20}];  
x2 = RandomReal[{5,10}];  
While[x1.x2 > 100, x1 = RandomReal[{10,20}]; x2 = RandomReal[{5,10}];]  
Return[x1.x2]
```

Figure 2.5 (right) shows the results of sampling from this model. The mean is around 82, the median is around 84, and its roughly 28%-likely to generate a number below 75.

So sampling is a flexible, general method for approximating arbitrarily-specified distributions. The catch? It’s *noisy*. Observing samples never removes all uncertainty about the distribution. Higher-order uncertainty should be expected.

But what, exactly, does a *higher-order uncertain* generative model—one that has uncertain opinions about its own sampling propensities—look like? As we saw above (§2.1.1), we want a probability function that assigns positive probability to assigning different values than it actually does. Understanding probabilities as sampling propensities, that means a generative model that has nontrivial sampling propensities for what its own sampling propensities are.

To see what this looks like, start by asking what it would mean for *Stan’s* generative model to be uncertain about *your* generative model’s sampling propensities. Let’s suppose that, like you, Stan is unsure whether or not I own dozen spoons—he’s 55%-confident that I do. Let’s suppose he’s sure that *you* are either 55%- or 45%-confident that I own a dozen spoons (i.e. that your model outputs d -samples either 55% or 45% of the time). Using ‘ P ’ as a description of your sampling propensities, let’s suppose he’s 60%-confident that $\langle P(d) = 0.55 \rangle$, and 40% confident that $\langle P(d) = 0.45 \rangle$. How

can his generative model encode these opinions? By generating samples from the product of the two questions, encoding a ‘joint distribution’ over them. In Stan’s case, here’s one way his joint distribution could be arranged to give the above verdicts. Suppose S_a , Stan’s actual sampling propensities, match those of an urn with the following numbers of 4 different types of marbles:

| | | |
|---|-----|-----------|
| | d | \bar{d} |
| $S_a = \frac{\langle P(d) = 0.55 \rangle}{\langle P(d) = 0.45 \rangle}$ | 45 | 15 |
| | 10 | 30 |

Summing the first column, Stan is $\frac{45+10}{45+10+15+30} = \frac{55}{100} = 55\%$ -confident that I own a dozen spoons. Summing the first row, he’s $\frac{45+15}{100} = 60\%$ -confident that *you* are 55%-confident that I own a dozen spoons. To model Stan as uncertain about your credences, we specify hypotheses about what your credences might be, and let his generative model have nontrivial sampling propensities over them.

What if Stan *elicits* your opinions by asking you to guess whether or not I own a dozen spoons? Suppose he knows you’ll draw 11 samples, and then guess d iff at least 6 of them are d -samples. Then he knows that the likelihood of your samples follow the distributions in Figure 2.6: if $\langle P(d) = 0.55 \rangle$, you’re 63%-likely to guess d , and if $\langle P(d) = 0.45 \rangle$, you’re 37%-likely to guess d . Thus if you guess d , that provides inconclusive evidence—Stan’s credence that your credence was 0.55 jumps from 0.6 to 0.72.¹⁵ Noise prevents him from being sure what you credence was.

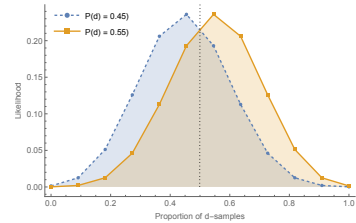


Figure 2.6: Propensities of Binomial(11,0.45) and Binomial(11,0.55).

Now, what does it look like for *your* generative model to be uncertain about *your* generative model’s sampling propensities? The same. At the actual world, you might even have the *same* generative model P_a that Stan does:

| | | |
|---|-----|----------|
| | d | $\neg d$ |
| $P_a = \frac{\langle P(d) = 0.55 \rangle}{\langle P(d) = 0.45 \rangle}$ | 45 | 15 |
| | 10 | 30 |

If so, your sampling propensities match the credences had in the actual world in the frame from Figure 2.1. You are in fact 55%-confident that I own a dozen spoons, are in fact 60%-confident that you are 55%-confident that I own a dozen spoons, and so on. And even if you elicit your own credences, noise implies that higher-order uncertainty remains.¹⁶

2.3 Noisy Sampling *Requires* Higher-Order Uncertainty

More is true. The sampling model *requires* higher-order uncertainty to generate noise. I’ll now argue that—at least for agents with opinions about their own opinions—reasonable sampling generates cognitive noise *only if* you have higher-order uncertainty.

¹⁵ $S_a(P(d) = 0.55|d) = \frac{S_a(P(d)=0.55) \cdot S_a('d'|P(d)=0.55)}{S_a(P(d)=0.55) \cdot S_a('d'|P(d)=0.55) + S_a(P(d)=0.45) \cdot S_a('d'|P(d)=0.45)} \approx \frac{0.6(0.63)}{0.6(0.63) + 0.4(0.37)} = 0.72.$

¹⁶If you know that you’ve conditioned on your elicitation, that changes your opinions: if P^+ is your posterior, you know that $P^+ = P(\cdot|elicitation = \epsilon)$. But since P^+ depends on what your prior (P) was—and you’re still unsure about that—you’ll still be unsure what P^+ is: P^+ will have higher-order uncertainty about P^+ . See Ch. 5.

This matters for two reasons. First, it contravenes standard uses of the sampling model in cognitive science. It's standard to assume two things. First, that the agent's underlying opinions evolve in a Standard Bayesian way: they start with a clear prior, and condition on a (partitional) signal. Second, the agent's noisy elicitations result from sampling from this posterior.¹⁷ This combination is a mistake. Though rarely realized, the first assumption implies that the agent's underlying opinions are *clear*, i.e. higher-order certain (see Chapter 3). Given this, the second assumption is sub-optimal: if such agent's can sample from their underlying distributions, then there's a *noiseless* way to elicit their opinions. If noisy-sampling is reasonable, our underlying opinions must be higher-order uncertain. But that, in turn, means that their opinions need not evolve in the way that the first assumption implies—as we'll see throughout this book, updating needn't be expected to improve accuracy or lead to convergence.

Second, and relatedly, my argument suggests that the sampling model has been misconstrued even by its proponents. They often say (using my terminology) that sampling can only explain *ambiguous* judgments—and that some other model is needed for *clear* judgments about fair coins and the like (e.g. Icard 2016, 891). I think this is also mistaken: once we pay attention to higher-order opinions, the sampling model *can* explain the continuum between ambiguity and clarity.

Return to our contrast. Let p be the claim that the 99th digit of pi is between 1–6, and let d be the claim that I own a dozen spoons. Your judgment about d is ambiguous and noisy. But your judgment about p is not: if I ask you on different occasions, you'll say '60%' every time. More generally: with clear judgments, there'll be little or no noise in your answers.

How can the sampling model explain this? It's not obvious. After all: on this model, to be 60%-confident of p is to have a generative model that outputs a p -sample 60% of the time. Won't *any* such model be noisy? Consider a computer program that draws a random integer between 1 and 100, and outputs ' p ' if it's between 1–60 and $\neg p$ if it's between 61–100. If we draw samples, there'll inevitably be variance in the proportion that are p -possibilities: even with 1000 samples, the proportion of p -samples will be above 61% or below 59% more than half time. Yet if I elicit *your* opinions about p , you'll say '60%' *more* than half the time. And if I ask you how likely a fair coin is to land heads, it's not as if you have some risk of blurting out '51%'. The sampling model's ability to explain ambiguous judgments seems to render it inappropriate for clear ones.

What to do? We might say that a different process governs clear judgments—that we just *know* certain probabilities. Of course, we *do* just know certain probabilities; but, because of this, sampling can explain clear judgments.

In fact, we need to account for a *continuum* of cases that hold fixed your credence in q , but vary how noisy your elicitations are. Recall Maggie: you were unsure how confident you are that she'll find a word in 7 seconds. Suppose she's done 100 word-completion tasks. How likely do you think she is to have found a word for a randomly-selected one? Suppose 50%. Elicitations of this credence will be noisy. But how noisy? It depends on what else you know. Consider:

Maggie-0: I tell you nothing else.

Maggie-1: I tell you she found words on between 1–99 of them.

¹⁷E.g. Gershman et al. 2012; Battaglia et al. 2013; Vul et al. 2014; Sanborn and Chater 2016; Zhu et al. 2020, 2023. Some ways of using sampling don't assume that the agent can update a clear prior on a signal (e.g. Ullman et al. 2014); they may avoid this objection.

Add a discussion of MCMP in computers; why can't we noiselessly sample from THEM? They are implicitly higher-order certain, but don't explicitly represent hypotheses about their own opinions. Could they? Logical omniscience is clearly an issue here...

...

Maggie-49: I tell you she found words on between 49–51 of them.

Maggie-50: I tell you she found words on exactly 50 of them.

It might well be that, in each case, your credence she completed a randomly-selected task is 50%. Yet the noise in your answers will decrease smoothly: in the *Maggie-0* case, you might well announce 70% or 30%; in the *Maggie-40* case, you’ll definitely announce between 40–60%; and in the *Maggie-50* case, you’ll always say 50%. Don’t take my word for it—an experiment (reported on below) confirms this observation. There’s a continuum between ambiguity and clarity.

And this continuum falls out of the sampling model. The trick? Higher-order opinions. Consider what the urn in your head looks like regarding both the 99th digit of pi and your opinion about this fact. You’re sure that you’re 60%-confident of this: $P(P(p) = 0.6) = 1$. Thus *every* sample you draw is one on which $\langle P(p) = 0.6 \rangle$, and 60% of them are ones in which p is true. The urn in your head looks like this:

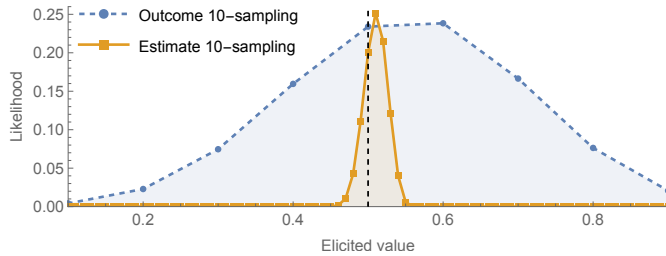
$$P_a = \begin{array}{c|cc} & p & \neg p \\ \hline \dots & \dots & \dots \\ \hline \langle P(p) = 0.7 \rangle & 0 & 0 \\ \hline \langle P(p) = 0.6 \rangle & 60 & 40 \\ \hline \langle P(p) = 0.5 \rangle & 0 & 0 \\ \hline \dots & \dots & \dots \end{array}$$

What happens if you draw samples? Although they’ll vary in the truth-value of p , *every* sample will be one on which $\langle P(p) = 0.6 \rangle$. Suppose you’re asked, ‘How likely do you think p is?’ You need to use these samples to determine your response. You *could* be flat-footed, announcing the proportion in which p is true—call that **outcome-sampling**, wherein your elicited probability of p is the proportion of samples in which p is true. That would induce noise.

But you shouldn’t be flat-footed. The question—‘How likely do you think p is?’—is asking about $P(p)$. Since you have higher-order opinions, that’s a quantity that your samples take a stand on—and *there’s no variance in it*. Every sample you draw is one in which the answer is ‘0.6’. So if you average the values of $P(p)$ in your samples, then—no matter what samples you draw—you’ll always answer ‘60%’. Likewise with the fair coin: since you know your credence in heads is 50%, averaging $P(h)$ in your samples always yields 50%. Instead of outcome-sampling, you can **estimate-sample**: let your elicited probability of p be the average value of *your credence in* p , $P(p)$, in the samples you drew. Indeed, under clarity estimate-sampling will always have higher expected-accuracy (relative to P) than outcome-sampling, since it’ll reduce random noise.¹⁸

Contrast ambiguous judgments. Recall the urn that modeled your (higher-order) uncertainty about whether I own a dozen spoons:

¹⁸Suppose you draw 10 samples from P_a and obtain the following elicitation: $\epsilon = \left(\begin{array}{ccc} \dots & p \& (P(p) = 0.6) & \neg p \& (P(p) = 0.6) & \dots \\ 0 & 5/10 & 5/10 & 0 & \end{array} \right)$. If you report your credence in p by outcome-sampling, you’ll say ‘50%’; but if you estimate-sample, you’ll say ‘60%’. The former is ϵ ’s expectation of p ’s truth-value, $\mathbb{E}_\epsilon(\mathbb{1}_p)$; the latter answer is ϵ ’s expectation of P ’s credence in p , $\mathbb{E}_\epsilon(P(p))$ —see Chapter 5.

$$P_a = \begin{array}{c|cc} & d & \neg d \\ \hline \langle P(d) = 0.55 \rangle & 45 & 15 \\ \hline \langle P(d) = 0.45 \rangle & 10 & 30 \end{array}$$


Any way of sampling from this model will be noisy: samples will vary in both the truth-value of d and the value of $P(d)$. If you draw 11 outcome-samples, the likelihoods be our familiar Binomial(11,0.55). And, if you draw 11 estimate-samples, tracking the *average value of $P(d)$* —the likelihoods will be the narrower (but still noisy) distribution plotted in orange.¹⁹ Estimate-sampling incorporates what you know about your own opinions: since you’re sure that your credence is between 45–55%, estimate-sampling always yields a value between these extremes.

I’m not a sampling imperialist. We know some probabilities without being able to simulate their domains. If scientists tell you there’s an 80%-chance dinosaurs were killed by a meteor, then you know that you think it’s 80%-likely without having any generative model of dinosaurs or meteors. But even here, you plausibly *can* simulate *your own beliefs* about dinosaurs—and since your judgment is clear, every sample will yield one in which $\langle P(\text{meteor}) = 0.8 \rangle$. This (higher-order) sampling model provides a general-purpose model of the contrast between ambiguity and clarity.

2.3.1 The Experiment

I’ve claimed that there’s a continuum between ambiguity and clarity: as you know more about the bounds of an (otherwise ambiguous) estimate, your elicitations will get less noisy. I ran simulations and an experiment to verify that ambiguous Bayesians and real people exhibit this pattern. Details are in the Appendices of this chapter; here are the takeaways.

The top row of Figure 2.7 shows the simulation results. I generated random probability frames,²⁰ laid a random variable X (with values between 0 and 1) over it, picked a random world, and gave the probability function at that world different levels of information about X . The ‘Lo Clarity’ condition learned nothing; the ‘Med Clarity’ condition learned that $\langle 0.25 \leq X \leq 0.75 \rangle$, and the ‘Hi Clarity’ condition learned that $\langle 0.46 \leq X \leq 0.54 \rangle$. I then elicited 4 (single) estimate-samples for X from the updated probability function, simulating eliciting the subject’s estimate for X on four separate occasions. The top left of Figure 2.7 shows a density plot of all elicitations across each condition—as you can see, across subjects the amount of variation in estimates is substantially reduced as clarity increases. This also applies to *within-subject* variation: the top right of Figure 2.7 shows density plots of within-subject standard deviations—how much each agent’s repeated elicitations tended to differ from each other. Increasing clarity decreases within-subject noise in estimates. This isn’t surprising: we’re just simulating agents who behave the way I described above.

¹⁹Here there’s less noise when estimate-sampling than outcome sampling; but there won’t always be. Estimate-sampling noise varies with how higher-order uncertain you are—it’ll steadily decrease from *Maggie-0*, to *Maggie-50*. And as this plot hints, under ambiguity estimate-sampling is often a *biased* indicator of your credences.

²⁰The frames were ‘factorable’ and hence Ambiguous-Bayesian, as discussed in Chapter 3.

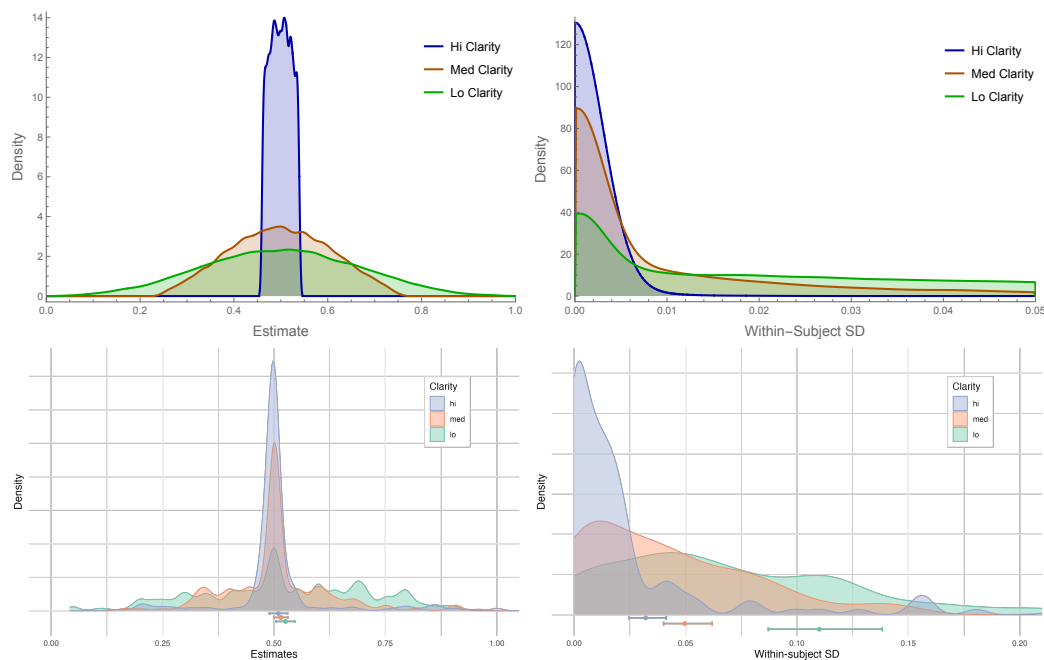


Figure 2.7: **Top:** Simulation results showing (left) individual estimates and (right) within-subject standard deviation in their estimates, divided by clarity. **Bottom:** Experimental results showing the same thing, along with Bayesian-regression estimates and 95%-credible intervals for the means below the x -axis.

A more interesting question is whether *real* people behave the same way. So I ran an experiment.²¹ The goal was to track intra-personal noise in estimates under ambiguity, varying how clear their estimates were in the way described with the *Maggie-0, ..., Maggie-50* examples. Subjects were told about word-completion tasks and given several examples. They were then presented with 8 different scenarios in which they were told that a new character was presented with n different completable word-search tasks (varying n), and were given limited information about how many they’d completed. Subjects were then asked to estimate exactly how many tasks the character completed.

They were randomized between Lo, Med, and Hi Clarity conditions. There were 4 target scenarios in which different conditions received different boundary information. The Lo condition was given no information about how many tasks the character completed (‘We haven’t started tallying her results yet’); the Med condition was told that they successfully completed at least 25% and at most 75% of the tasks; and the Hi condition was told that they completed at least 46% and at most 54% of the tasks. To force subjects to re-estimate each time, the 4 scenarios varied the total number of tasks ($n = 100, 126, 72, 44$)—so, for instance, for the 44-task scenario the Med subjects

²¹Pre-registration here: <https://aspredicted.org/2swj-sj45.pdf>. I recruited English-speaking UK residents on Prolific. Subjects were initially given a comprehension check on estimates, and automatically screened out if they failed. With a target of 400, Prolific ended up recruiting 444 subjects, 399 (89%) of which passed this check. An additional 29 subjects failed an attention check later in the study and were also excluded, leading to an overall pass-rate of 83%.

were told, ‘We haven’t finished tallying her results yet, but we know she successfully completed at least 11 and at most 33 of the tasks’, whereas in the 100-task they were told ‘at least 25 and at most 75’. In addition to the 4 target scenarios, 4 filler scenarios were included to prevent subjects from noticing patterns in the information they were given.

The first main prediction was that subject’s mean estimates would not differ substantially across conditions, even as estimate variance did. I normalized each subject’s 4 target estimates to 0–1, averaged these to get their within-subject mean estimate, and ran a Bayesian Gaussian regression in BRMS to estimate the average subject’s response in each condition. (We’ll talk more about how to interpret statistics like this in §3.4. For now, just think of the estimates as saying what a simple-minded Bayesian agent who’s perfectly incorporated the experimental data believes about the averages in the different groups.) The model estimates are plotted with 95%-credible intervals below the x -axis in the bottom left of Figure 2.7; as you can see, all were quite close to each other, slightly about 50%. This plot also shows density plots of all subjects’ (normalized) estimates, showing that variance goes down as clarity increases.

But this doesn’t show that *intra*-personal noise decreases. To do that, I calculated within-subject standard deviations amongst their 4 normalized estimates, and then ran a hurdle log-normal regression in BRMS—a standard distribution for fitting non-negative (but sometimes 0), right-skewed outcomes, like our within-subject standard deviations. The mean estimates and 95%-credible intervals are plotted below the x -axis on the bottom right of Figure 2.7, along with density plots of the raw data. As predicted, within-subject noise in estimates decreased markedly as clarity increased.

Upshot: in theory and in practice, reducing ambiguity reduces noise in estimates.

2.3.2 The Argument

Let’s turn that theory into a precise argument. The claim: on the sampling model, if cognitive noise is inevitable for an agent, then they must have higher-order uncertainty. I’ll show the contrapositive: if the agent has higher-order *certainty*, noise is avoidable.

Assume the sampling model—to have credence y in p is for your generative model to have y -chance of generating a p -sample:

(P1) Sampling. If $P(p) = y$, then each sample you generate is y -likely to be one in which p .

The second premise is that, when asked to estimate the value of a quantity X , an available strategy is to give X ’s average value amongst the samples you drew:

(P2) Mean elicitation. An available strategy for estimating the value of X is to elicit samples and then report the mean value of X amongst them.

Notice that if X is the truth-value of q — $\mathbb{1}_q$ —this is outcome-sampling, while if X is your credence in q — $P(q)$ —this is estimate-sampling.

Now suppose that you have clarity: there is an x such that $P_a(P(q) = x) = 1$. From P2, one strategy is to estimate-sample, eliciting samples and then give the average value of $P(q)$. By hypothesis, you are certain that $\langle P(q) = x \rangle$. From P1, it thereby follows that each sample you

Should this come before the experiment?

draw has a 100%-chance of being one in which $\langle P(q) = x \rangle$. So if you report the mean value of $P(q)$ in your samples, you'll always (noiselessly) report x .

P1 defines the sampling model, while P2 captures how samples influence your behavior. Neither can be rejected by those who accept the sampling model. So on the sampling model, whenever you have higher-order *certainty*, noise is avoidable. Contraposing: on the sampling model, whenever noise is *inevitable*, you must have higher-order *uncertainty*. As mentioned, this contravenes many uses of sampling, so let me address some objections.

Objection: What about generative models that have no opinions about their own opinions?

First reply: we *do* have opinions about our own opinions. So if the sampling-model applies to *credences*, its models must include such higher-order opinions.

Second reply: You didn't have a model my spoon collection ready to go; you had to construct it on the fly. (Some argue that this on-the-fly model construction is key to general intelligence; see Icard and Goodman 2015; Brooke-Wilson 2024.) *Maybe* we don't have models of our own opinions ready-to-go. (Though I bet we do—many actions, like setting an alarm or writing a to-do list, rely on beliefs about your own future- or current-beliefs.) But the argument still shows that *once we construct them*, higher-order certainty would make noise disappear. Since it doesn't, the models we construct must have higher-order uncertainty.

Objection: The sampling model implies that *ideal* versions of ourselves would have clarity. For if we sample *enough*, we can learn what our credences are.

Reply: Granted, on this simple model you can figure out (roughly) what your opinions are by sampling many times and updating on the results. But, under higher-order uncertainty, learning what your opinions are would *change* your opinions (see §3.1.2). So the opinions you'd have in the limit of sampling *and updating on what those samples were* are not your current subjective probabilities opinions—they are your 'informed probabilities' in the terminology of Chapter 3. Your current probabilities are the dispositions you have now, reflected your current sampling propensities—not the dispositions you'd have after reflecting on data that you yourself have generated.

Consider an analogy. You can approximate pi by expanding the Nilakantha series (Irkhin 2022): $3 + \sum_{n=1}^{\infty} \frac{4(-1)^{n+1}}{(2n+1)^3 - (2n+1)}$. So you currently have the following disposition: if you were given enough time and incentive, you'd respond with certainty about the 99th digit of pi. This doesn't show that your 'true' opinion is certain of whether the 99th digit is between 1–6. You truly *are* 60%-confident. Focusing on (noisy approximations of) the idealized version of yourself that's done these calculations would make bad empirical predictions and incorrect normative verdicts. You have no chance of betting the farm against p , nor are you irrational for failing to do so.

To vary the example (and avoid concerns about logical-omniscience—see footnote 14) it may well be that if last night I'd had unlimited time and paper for journaling, I could've worked out whether I was stressed or dehydrated. It doesn't follow that my 'true' opinion was certain of whether I was stressed or dehydrated. I truly was uncertain.

My claim is that the analogy is apt: focusing on the idealized versions of people who are higher-order certain distorts both our empirical predictions and normative verdicts.

Objection (based on MCMF): Argument presupposes logical omniscience? Maybe $P(P(q) = \pi(q|data))$ is 1, but uncertain what $\pi(q|data)$ is. Of course, that's a way of being unsure what your prior P is. Hard interpretative Qs here...

2.4 Cognitive Noise Implies Higher-Order Uncertainty

Once we've seen the sampling model, it's easy to see how other models of credences can allow higher-order uncertainty. All agree that having a credence of 0.6 in q is a matter of having the right dispositions, like accepting bets on q . But we're often unsure what our dispositions are—and since they're noisy, our actions provide inconclusive evidence about them. Maybe one way to have credence 0.6 in q is to be disposed to draw a number x from a (say) Normal distribution with mean 0.6, and then use x to make your decisions (cf. Erev et al. 1994). Clearly this will lead to higher-order uncertainty—and it's easy to imagine variants.

Abstracting over such alternatives, I'll now give two arguments that—in general—agents with cognitive noise should be expected to have higher-order uncertainty.

The first argument takes four sentences. Timothy Williamson has (in)famously argued that for agents with cognitive noise, there is *no* nontrivial condition such that we're always in a position to know whether or not it obtains (Williamson 2000, 2008; Srinivasan 2015a). *You have credence 0.6 that I own a dozen spoons* is a nontrivial condition. So if Williamson's right, you can't always know whether or not you have credence 0.6 that I own a dozen spoons. Since you sometimes *know* that you can't know this—and you shouldn't be certain of things you know you can't know—you should sometimes be uncertain what subjective probabilities you have.

The second argument is a bit longer, but the idea is simple: if you were certain of what your credences are, then you'd be able to reliably *say* what they are. But you can't. So you aren't.

Consider a paradigm case of ambiguous judgment. Assume you're reasonable, and that conditions are normal—no hypnotists, guns to the head, etc. How likely do you think it is that I own a dozen spoons (d)? Your judgment is noisy:

(P1) Cognitive Noise. In estimating your credence in d under ambiguity, if you're x -confident, there's some chance you'll estimate a number higher or lower than x .

Precisely: if $P(d) = x$, then there's an $y > 0$ such that there's a decent (say, >20%) chance that you'll give an estimate for $P(d)$ that's higher than $x + y$ or lower than $x - y$.

This premise only applies to ambiguous judgments in normal circumstances. The 20% threshold is arbitrary—what we need is that as y goes to 0, the chance of mis-estimating by at least y will rise above any reasonable threshold.²² To deny P1 would amount to saying you're infinitely reliable in your estimates of your own opinions. We can imagine agents whose elicitations of their mental states are that reliable. They are not us.

The second premise links your certainties to your elicited estimates (in normal conditions):

(P2) Certainty Precludes Chance. For any X , if you're certain that X is between x_1 and x_2 (i.e. $P(x_1 < X < x_2) = 1$), then you can elicit an estimate for it in a way that has a high (say, $\geq 90\%$) chance of being between x_1 and x_2 .

The 90%-threshold is again arbitrary—it's just there to show that we needn't assume that your certainties *perfectly*-reliably constrain your estimates. To deny P2 would be to adopt an anti-dispositionalist understanding of subjective probabilities: what do credences *do* if being certain

²²If you round your estimates, understand P1 as forcing you to expand your estimate to arbitrary precision.

that a quantity is within a range doesn't constrain your estimates of it to that range?²³

P1 and P2 together imply that you have higher-order uncertainty. Suppose—for reductio—that you were certain what your credence in d was: there is an x such that $P(P(d) = x) = 1$. By P1, there's a y —say, $y = 0.001$ —such that you have more than a 20%-chance of giving an estimate that's below $x - 0.001$ or above $x + 0.001$. Hence you have less than an 80%-chance of giving an estimate between $x - 0.001$ and $x + 0.001$. By supposition, you are certain that $(x - 0.001 < P(d) < x + 0.001)$. But P2 then implies that you have at least a 90%-chance of estimating $P(d)$ between $x - 0.001$ and $x + 0.001$. That's a contradiction, so we reject our supposition: when P1 and P2 hold, your credence in d must be higher-order uncertain.²⁴

Upshot: in general, if a reasonable agent inevitably has cognitive noise when they try to elicit their opinions, then they must have higher-order uncertainty.

2.5 Where Clarity Hides in Cognitive Science[†]

Cognitive scientists may be puzzled. I've argued that their models often make two assumptions that are in tension: (1) that agents start with a clear prior and update it by conditioning it on the value of a signal, and (2) that elicitation of their opinions are noisy. If (1) is true, then (2) is suboptimal. Chapter 3 will explain how clarity about the prior slides in, but it's worth working through an example now.

First, we need to distinguish *noisy-signal* from *noisy-elicitation* models. Noisy-signal models assume that you start with a prior distribution P_a that's clear (since it's not modeled as a variable), and that at any given moment you update P_a by conditioning it on some noisy signal (see Woodford 2020). The idea comes from perception: when you look at a tree and try to estimate its height, there's randomness in the visual signal (Green and Swets 1966). But it's often adapted to internal signals, too: when you think about how many spoons I might own, it's noisy what you call to mind to elicit your opinion (e.g. Enke and Graeber 2023—see §3.3).

Here's a standard example. You're staring at a tree some distance off, trying to estimate its height, H . You (know that you) will get a 'height impression' signal, S that's normally distributed around the tree's true height (say, with known variance—but this is inessential; what's essential is that you are certain of what your likelihoods $P(S = y|H = x)$ are). Then if you start with a clear prior P_a that's uniform over the tree being between 10 and 30 meters tall, and condition it on your observed height impression, your posterior over the height will be normally distributed with mean equal to your observed signal. This is illustrated in Figure 2.8: the x -axis is the tree's actual height

²³Outcome-sampling can lead to violations of Certainty Preclude Chance, but estimate-sampling never will. And P2 only says you *can* elicit an estimate with the relevant reliability.

²⁴It might be tempting to reach for imprecision, denying that there's a number x such that $P(d) = x$. Perhaps there's merely a *range* of credences $[l, h]$ such that your credence in d is determinately at least l and determinately less than h , but it's indeterminate beyond that. This doesn't help. If our credences are imprecise, then there are facts about *how* imprecise: instead of a real-valued variable $P(d)$ from possibilities w to numbers $P_w(d)$, we have a set-valued variable $\mathbb{P}(d)$ from possibilities w to sets of numbers $\mathbb{P}_w(d)$. And just as noise prevents us from being sure of which exact number x is our credence in d (assuming we have one), so too noise prevents us from being sure of which exact interval $[l, h]$ is our imprecise credence in d (assuming we have one). If you're convinced we have imprecise credences, then noise shows that we have *both* credal imprecision *and* higher-order uncertainty (cf. Molinari 2023). The former can't play the role of the latter.

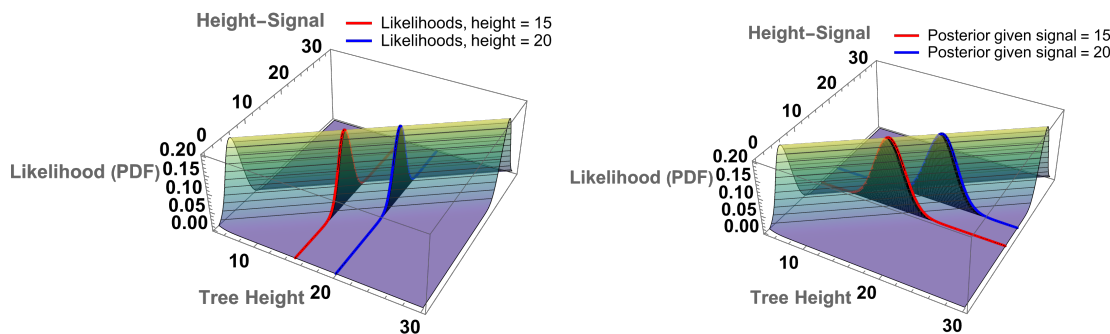


Figure 2.8: Your prior over the tree’s height and the height-signal you’ll receive, assuming the signals are normally distributed around the true height. The red and blue curves on the left show the likelihood distributions for the height-signal, conditional on the true height being 15 and 20, respectively. The red and blue distributions on the right are the posteriors you’d have after conditioning on the height-signal being either 15 or 20, respectively. Note that they assign each other 0 probability—they are certain which signal they received.

(in meters), the y -axis is the height signal, and the z -axis (coming out of the page) is your clear prior distribution over the two. In the left plot, the red and blue curves represent the likelihoods of getting various height signals if the tree’s true height is 15 or 20 meters, i.e. $P(\cdot|H = 15)$ or $P(\cdot|H = 20)$. In the right plot, the red and blue curves are different possible posteriors, conditioning your prior on the height-signal being 15 or 20, i.e. $P(\cdot|S = 15)$ or $P(\cdot|S = 20)$.

In models like this, the state of the external system (the tree’s height) doesn’t determine your reported posterior. But state of the world does, since *which signal you received* is part of the world, which you (beforehand) can be uncertain about (contra Isaacs and Russell 2022; Easwaran and Nielsen 2023). And the way it does so guarantees that your posterior is clear, for two reasons. First, your prior is clear—there is no uncertainty over what initial distribution you have over the joint height-and-signal state space. (Mathematically, this is smuggled in by making your prior P_a a constant, rather than a random variable—see Ch. 3.) Second, since the signals are partitional, your posterior always assigns probability 0 to receiving any signal other than the one it actually received. Notice in the right diagram that although the red and blue posterior distributions overlap in their distribution over the tree’s height (H), they assign 0 probability to each other’s signal (S). Since your posterior is certain of what your prior was and what signal it received, it has clarity.

There is plenty to like about these models. And if they say that your elicited posterior is always equal to your prior conditioned on the signal, then they don’t face the main objection I’ve raised in this chapter—for they don’t say that there’s noise in the *elicitation* of your credence, but in its fluctuations. Of course, they face other objections. They are Standard-Bayesian models, and therefore—unless supplemented with lossy memory or other foibles—satisfy Reasonable Convergence (and so are empirically falsified). And they don’t explain *how* your prior is clear: since your current credence is constantly fluctuating, observing your behavior is no reliable guide to what you prior was. They would be more plausible and explanatorily-powerful if they permitted ambiguity.

But the real trouble comes at the next step. *Noisy-elicitation* models (e.g. those in fn. 17) further assert that rather than eliciting your posterior distribution directly, your elicitation is the result of noisily *sampling* from your posterior. For example, they say that if you’ve learned

that $S = 20$ and are asked how likely it is that the tree is over 21 meters tall ($H > 21$), you draw samples from $P(\cdot|S = 20)$ and report the proportion in which $H > 21$.

This won't do. If the model of your underlying opinions is right, then your posterior is clear and assigns probability 1 to having the distribution it actually has— $P^+(P^+(\cdot) = P_a(\cdot|S = 20)) = 1$. This posterior assigns 0.3085 credence to $\langle H > 21 \rangle$, so *any* sample drawn from this distribution will be one in which $\langle P^+(H > 21) = 0.3085 \rangle$; drawing a single sample will suffice to reliably report it.

Obviously people can't do that. But if their priors were clear, they could. Thus we've misrepresented their prior uncertainty by implicitly assuming that they were certain that their prior was the distribution P_a . Realistically, your prior was uncertain about exactly what its distribution over the height of the tree was, as well as its likelihoods for receiving a given signal given that the tree was a given height: P_a is unsure both of what $P(H = 21)$ and what $P(S = 20|H = 21)$ are.²⁵

This is not just a 'philosophical' objection. Moving to Ambiguous-Bayesian models not only gives a stronger normative justification for noisy sampling, but also fundamentally changes the dynamics of belief-updating. Standard Bayesians will never exhibit hindsight bias (Ch. 7), predictably shift their beliefs in the direction they are searching for evidence (Ch. 8), or predictably polarize in the face of plentiful evidence (Ch. 9). Real people do. We cannot explain or rationalize such behavior by laying noise on top of Standard-Bayesian models. But I'll show that if we expand our models to allow ambiguity in the prior, we can. Even on their own terms, cognitive scientists should take seriously ambiguity in *beliefs*—not just in elicitations of those beliefs. Their theories will be more explanatory and predictive if they do.

2.6 Upshots

If we use probability(-like²⁶) judgments to manage uncertainty—as pretty much every serious cognitive-scientific theory says we do—then cognitive noise leads to higher-order uncertainty. So all should agree that we have not merely *probabilistic* uncertainty—uncertainty about what a more-ideal versions of our probabilities is—but *higher-order* uncertainty: uncertainty about what our own subjective probabilities are. That makes higher-order uncertainty the natural way to model ambiguity. Yet few theories make use of it—pretty much every Bayesian theory (implicitly) presupposes higher-order certainty (see Chs. 3–4).

Why? The difficulty is that higher-order uncertainty is a quagmire of mathematical and conceptual subtlety. The next chapter gives a non-technical introduction to the key concepts and distinctions needed to think clearly about it. Read that, and then decide whether you want more details about the foundations of higher-order uncertainty (Part II), or want to jump to its applications (Part III).

²⁵And perhaps your posterior is even uncertain of which signal it received—see 4.4.1 for discussion.

²⁶Notice: the same arguments for higher-order uncertainty would work for non-probabilistic measures of degrees of confidence (Ch. 1, footnote 3).

2.7 Appendix: Theoretical Details[†]

The full code for replicating the simulations is in the companion Mathematica notebook.²⁷ The design was simple. I generated random, regular (all worlds assign each other positive probability) Ambiguous-Bayesian frames with 10 worlds each (see Chapter 3). A world is picked uniformly at random to be the actual world. Then a variable X is constructed at random by assigning each world an independent, random draw from a Beta(0.5,0.5) distribution—so it is bounded between 0 and 1 but has a fair bit of variance. The relevant claim about its value is then constructed by picking the worlds at which it's true: $\langle 0 \leq X \leq 1 \rangle$ for the Lo Clarity condition; $\langle 0.25 \leq X \leq 0.75 \rangle$ for the Med Clarity condition; and $\langle 0.46 \leq X \leq 0.54 \rangle$ for the Hi Clarity condition. Every world in the frame is conditioned on this relevant information, and then 4 (single) estimate-samples of X are drawn from the updated actual distribution.

For each condition, I recorded (1) the estimate samples drawn, (2) the within-subject means of their estimate samples, and (3) the within-subject standard deviation of their estimate samples. Data from (1) and (3) were used to generate the density plots shown in the top row of Figure 2.7.

2.8 Appendix: Empirical Details[€]

2.8.1 Data collection

The experiment aimed to recruit 400 English-speaking UK residents after the initial comprehension check. Prolific ended up recruiting 399 participants whose data was recorded, out of 444 who started the experiment. Pre-registration here: <https://aspredicted.org/2swj-sj45.pdf>. The full Qualtrics survey as well as the raw data and R script are available in the ResearchBox.²⁸

The comprehension check was designed to prompt people to think of estimates as mathematical expectations—see the ResearchBox for details.

Subjects were then told about (completable) word completion tasks, and given three examples along with their completions: P_A_ET (planet), B_L_E (belie, belle, bulge), and TH__N (thorn). They were randomized into Lo, Med, and Hi Clarity. Each condition saw, in randomized order, 8 estimate scenarios and 1 attention check which instructed them on how to answer. The data from those (29 subjects) who failed the attention check were excluded from the study, as pre-registered. All in all, 370 of 444 (83%) of all initial recruits passed both the comprehension and attention checks. In addition, one participant entered a single estimate which was an order of magnitude above the upper bound (estimating that Fiona completed 363 of 72 word-completion tasks). This was probably a typo; all of these participants responses were post-hoc excluded to avoid the nonsensical outlier, leading to a total of 369 subjects.

The 8 estimate scenarios went as follows. Subjects were told about a character (all women with generic names, to emphasize that they were different characters but avoid effects of believed demographic differences in word-search ability) who had been presented with a certain number of word-completion tasks, were told limited information about how many they had completed, and

²⁷TODO [link to notebook]

²⁸TODO: [link to researchbox]

Table 2.1: (MAIN-1), Fixed effect estimates

| Intercept | mo(clarity) | Residual SD |
|-------------------|---------------------|-------------------|
| 0.53 [0.50, 0.55] | -0.01 [-0.02, 0.01] | 0.13 [0.12, 0.14] |

Brackets are 95%-credible intervals. Baseline clarity is Lo.

were then asked to estimate the exact number. The 4 target scenarios varied how much information was given, by clarity. The Lo Clarity condition was always given no information, while the Med Clarity was told that she completed between 25%–75% of the total tasks, and the Hi Clarity condition was told that she completed between 46%–54% of them. The total number of tasks varied, to force subjects to re-estimate each time. See the ResearchBox for screenshots.

The 4 remaining scenarios were fillers which had the same structure but whose data were not analyzed. These tasks gave intervals that were not centered on 50% (but were counter-balanced around it, across scenarios), in order to prevent subjects from noticing the pattern. All conditions saw the same fillers.

2.8.2 Analyses

Following the pre-registration plan, I first normalized subjects estimates by dividing by the total number of tasks—so, for instance, an estimate of 34 in response to Fiona’s scenario above would get converted to an estimate of $\frac{34}{72} = 0.472$. I then calculated each subjects’ normalized mean of their 4 target estimates, recorded as `norm_mean`, and their normalized within-subject standard deviation between their 4 target estimates, recorded as `norm_sd`.

Bayesian regressions were run in BRMS using default priors (flat over fixed effects), using the `seed=123` for reproducibility. Since all analyses have 1 datapoint per participant, I omitted random effects. In addition, I instructed BRMS to treat clarity as an ordered variable with a monotonic effect, using `mo(clarity)`—this allows it to try to estimate what proportion of the effect is due to the Lo→Med step and the Med→Hi step.

(Main-1). The first main prediction was that there would not be large differences between subjects’ `norm_mean` across conditions. I ran the following Gaussian Regression:

$$\text{norm_mean} \sim \text{mo}(\text{clarity})$$

Full regression outputs are in the ResearchBox. Fixed-effect estimates and 95%-credible intervals are in Table 2.1. As you can see, estimates are precise for a 0.53 mean estimate in the Lo condition, with a precise estimate of no large effect of clarity: the estimate is -0.01 , with a 95%-credible interval of $[-0.02, 0.01]$. This is the model’s estimate for the total effect of moving from Lo to Hi along the clarity scale. It’s estimates and 95%-credible intervals for the `norm_mean` in each condition are: Lo, 0.53 [0.50, 0.55]; Med, 0.52 [0.50, 0.53]; Hi, 0.51 [0.49, 0.53]. These model estimates for each condition were plotted under the x -axis in the bottom left plot of Figure 7, along with density plots of the normalized estimates in each condition (not `norm_mean`, which would show less variance).

(Main-2). The second main prediction was that increasing subjects’ clarity would decrease the

noise in their estimates, i.e. decrease `norm_sd`. I estimated this with a hurdle (i.e. zero-inflated) log-normal Bayesian regression in BRMS:

```
bf(norm_sd ~ mo(clarity)
  hu ~ mo(clarity))
```

This is a standard model for estimating the mean of a distribution that is non-negative, sometimes 0, and right-skewed—as our `norm_sd` is. Basically, the model tries to estimate both (1) how likely a given datapoint is to be 0 (it fails the ‘hurdle’, as happens when all 4 of a subject’s normalized estimates are identical), and (2) how large the datapoint is likely to be given that it’s nonzero. Part (2) uses clarity as a monotonic predictor of `norm_sd`, using log-normal likelihoods; part (1) uses clarity as a monotonic predictor of the probability that a data-point is 0, using Bernoulli likelihoods and a logit link.

Full regression outputs are in the ResearchBox. Fixed-effect estimates and 95%-credible intervals are in Table 2.2. Intercept and `mo(clarity)` are reported on a log scale, while the `hu-intercept` and `hu-mo(clarity)` terms are on a logit scale. These parameters are hard to interpret directly, but the two key facts are (1) the `mo(clarity)` term is significantly below 0, meaning that the nonzero part of the model predicts that as clarity goes from Lo to Med to Hi, the nonzero `norm_sd` go down; and (2) that the `hu-mo(clarity)` term is positive, meaning that the model predicts that moving from Lo to Med to Hi clarity makes it more likely that the `norm_sd` is 0. Inspecting the regression output reveals that it predicts that 81% of effect (1) happens on the Lo→Med step, with 19% in the Med→Hi step; while it predicts that 9% of effect (2) happens on the Lo→Med step, and 91% of it happens on the Med→Hi step. This makes sense, intuitively: it’s only when the boundary information gets narrow that it becomes particularly likely that there’s zero variance in estimates.

Table 2.2: (MAIN-2), Fixed effect estimates

| Intercept | mo(clarity) | hu-intercept | hu-mo(clarity) | Residual SD |
|----------------------|----------------------|----------------------|-------------------|-------------------|
| -2.81 [-3.02, -2.60] | -0.49 [-0.64, -0.34] | -3.05 [-3.73, -2.46] | 1.00 [0.63, 1.39] | 1.14 [1.05, 1.23] |

Brackets are 95%-credible intervals. Baseline clarity is Lo.

Reading this table directly: conditional on the data-point being positive, the model predicts that in the Lo condition the median `norm_sd` is $e^{-2.81} \approx 0.060$, that moving to the Med condition lowers it by $-0.49(0.81)$ to $e^{-3.21} \approx 0.040$, and moving to the Hi condition lowers it by an addition $-0.49(0.19)$ to $e^{-3.3} \approx .037$. Meanwhile, the hurdle part of the model predicts that the Lo condition has a $\text{Logistic}(-3.05) = \frac{1}{1+e^{3.05}} \approx 0.045$ chance of being a 0-value, that moving from Lo to Med increases this chance to $\text{Logistic}(-3.05 + 1.00(0.09)) \approx 0.049$, and moving from Med to Hi increases it to $\text{Logistic}(-3.05 + 1.00) \approx 0.114$.

Combining these components, overall the model predicts that the *mean* `norm_sd` of Lo-Clarity subjects is 0.11 [0.09, 0.14]; of Med-Clarity subjects is 0.05 [0.04, 0.06]; and of Hi-Clarity subjects is 0.03 [0.02, 0.04]. These model predictions are plotted below the x -axis in the bottom right plot of Figure 7, along with the raw `norm_sd` data.

(Supp-1). Finally, I pre-registered a supplemental analysis uses bootstrapping to estimate 95%-

confident intervals for the median `norm_sd` in each condition, as a sanity check on the Bayesian regression. This went as expected, noting that the median will be lower than the mean in a right-skewed distribution like this: the estimate and intervals for the median Lo `norm_sd` was 0.061 [0.049, 0.078], for Med was 0.035 [0.026, 0.045], and for Hi was 0.012 [0.009, 0.016].