Subjective Probability as Sampling Propensity*

(To appear in *Review of Philosophy and Psychology*)

Thomas F. Icard, III
Stanford University

Abstract

Subjective probability plays an increasingly important role in many fields concerned with human cognition and behavior. Yet there have been significant criticisms of the idea that probabilities could actually be represented in the mind. This paper presents and elaborates a view of subjective probability as a kind of *sampling propensity* associated with internally represented generative models. The resulting view answers to some of the most well known criticisms of subjective probability, and is also supported by empirical work in neuroscience and behavioral psychology. The repercussions of the view for how we conceive of many ordinary instances of subjective probability, and how it relates to more traditional conceptions of subjective probability, are discussed in some detail.

1 Introduction

Subjective probability is a ubiquitous tool in philosophy, psychology, the social sciences, and elsewhere, for understanding aspects of intelligent inference, reasoning, and decision making under uncertainty. The basic idea is that numerical probabilities, in the familiar mathematical sense, can be used to represent a person's subjective "strength" or "degree" of conviction or belief. There is no general consensus about what exactly these numbers mean, and how (if at all) they might be literally represented in an agent's mind, though there are some classic analyses of the *concept* of probability. Foundational discussions of subjective probability, beginning with Ramsey, de Finetti, and others, typically analyze or define the concept of subjective probability by reducing it to some observable or measurable phenomenon, such as preference or betting odds. As Ramsey famously put it, "a degree of belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it" (Ramsey, 1931, 71).

While these concepts of probability—tied intimately to behavioral dispositions of a whole person—may be useful for many purposes, they are not evidently appropriate to explain the role that subjective probability plays in contemporary psychology and cognitive science. In much of this work, especially recently, the proposal is not just that we can reasonably ascribe subjective probabilities to people, and thereby explain much of their behavior. There is a further claim that the brain somehow encodes probabilities and uses these in the service of prediction and action selection (for example, see Knill and Pouget 2004; Gopnik et al. 2004; Chater et al. 2006; Yang and Shadlen 2007; Vilares and Kording 2011; Perfors 2012; Clark 2013, among many others); that is, something like subjective probabilities are assumed to play a direct causal role in pervasive mental computations. While much of this work is about low-level cognition, especially perception, increasingly more work is aimed at ordinary inference and prediction about "everyday" events and topics (see, e.g., Griffiths and Tenenbaum 2006; Goodman et al. 2014, *inter alia*), thus coming quite close to the sorts of commonplace inferences and judgments that are the focus of many philosophers interested in subjective probability and its associated normative questions. Given the increasing prevalence of probability in computational modeling of cognition, it is important to understand both (1) what, concretely, the relevant states of mind are supposed to be, and (2) how the resulting notion of subjective probability relates to the more traditional conceptions.

In this paper we present, explain, elaborate, and aim to elucidate a recently proposed hypothesis about how probabilities might be concretely represented. *The Sampling Hypothesis* asserts that an important way in which the brain

*This is a preprint. The final version will appear in *Review of Philosophy and Psychology*. Thanks to the RoPP editor Paul Egré, to the journal reviewers, and to Falk Lieder for useful comments that helped improve the paper. Thanks also to Wesley Holliday, Shane Steinert-Threlkeld, and my dissertation committee (see Icard 2013) for helpful comments on an earlier version.

represents probabilistic information is via generative models whose underlying "sampling propensities" implicitly encode probabilities for possible states of the world. That is, a person's subjective probabilities for a given event are encoded roughly by the proportion of times the person's "intuitive model" of the world would be expected to *sample* that event from among the possible outputs of the model. In other words, we replace Ramsey's *propensity to act* with an internal *sampling propensity* of the person's intuitive model, effectively reducing subjective probabilities to a particular kind of (at least relatively) objective probability.

The Sampling Hypothesis has received some impressive empirical support over the last several years, and the cognitive and neuroscientific literature on this topic is complex and growing. This paper will not contain a comprehensive survey. Instead, the aim is to give a general sense of the empirical support for the hypothesis, to make a convincing case that there is something to the hypothesis, and to explore some of the consequences of the overall view for how we should think about subjective probability. Rather than replacing more traditional analyses of subjective probability, the intention is to understand an important *variety* of subjective probability that can play a different theoretical role, one more appropriate to its use in psychology, and perhaps also in much of philosophy, by capturing part of what guides our everyday inferences, predictions, judgments, and actions.

The outline of the paper is as follows. In §2 we offer some background discussion about the *concept* of subjective probability, what theoretical role it is supposed to play in the psychological and behavioral sciences and in philosophy, and what we take to be the conceptual core of subjective probability. In §3 and §4 we review commonly expressed doubts about the very idea that the mind could harbor numerical probabilistic representations. In §5 we introduce the proposal in detail (with some mathematical details left to Appendix A), and offer clarificatory remarks in §6 and §7. Important empirical evidence—centered around neural computation, heuristics and biases, and the phenomenon of probability matching—is reviewed and discussed in §8, §9, §10, and Appendix B. Challenges to the hypothesis are presented in §11, and a final comparison with the traditional conceptions of subjective probability is offered in §12.

Our goal is not to convince the reader that the Sampling Hypothesis must be correct, but rather to offer a proposal about what exactly the hypothesis comes to, and show why it is a promising, attractive idea worth taking seriously.

2 Analyzing Subjective Probability

The standard formalization of probability theory assumes a *sample space* $\mathcal{V}$, together with some non-trivial sigma-algebra $\mathcal{E}$ over $\mathcal{V}$, and a measurable function $P : \mathcal{E} \to [0, 1]$ satisfying the usual axioms of (at least finite) additivity and normalization (that $P(\mathcal{V}) = 1$). The framework can be analyzed and applied with the basic mathematical entities—in particular the probabilities over "events" in $\mathcal{E}$—interpreted in numerous ways. For instance, on a *frequency* interpretation, the probability values describe objective (limiting) frequencies of events in the world, relative to some background reference class. On a *propensity* interpretation, probabilities describe real dispositions or objective physical tendencies. One of the most important interpretations for the behavioral and social sciences is the *subjective* interpretation, according to which these probabilities represent a "degree of confidence" or some other measure of how likely things are from a specific agent's subjective point of view. The following are some examples of typical phenomena to which the probability calculus, under a subjective interpretation, has been applied:

1. A government must decide whether, and how much, to invest in R&D on alternative energy.
2. A doctor is trying to decide whether to prescribe rest or a penicillin shot for a patient with a soar throat.
3. Seeing a cloudy sky, a person must decide whether to walk home and risk rain, or take the bus.
4. An interlocutor must determine whether a speaker is talking about a person named 'Paris' or a city of that name.
5. A monkey must "decide" whether to saccade to the left or right, in order to receive a payoff.
6. The visual system must determine how far an object is, given a pattern of input on the retina.

This short list is already quite diverse. In scenarios like 1 and 2 (and perhaps 3), many claim that the probability calculus (together with a theory of utility, and perhaps other ingredients) plays into a compelling normative account of what one *ought* to do. After assessing the situation, figuring out what the possible actions and outcomes might be,

and assigning probabilities to different eventualities in a way consistent with the probability axioms, one ought assess actions by calculating their expected utilities and choosing one whose expected utility is non-dominated. In 4-6 (and perhaps 3), by contrast, the claim is typically that our brains somehow calculate probabilities "for us," and that the relevant computations are happening "under the hood," so to speak, not under conscious control. Despite this diversity, in all of these cases, the probabilities in question—whether concerning possible advances in energy development, what entity a speaker is likely to be referring to, or spatial distances—have been understood as subjective. In advance of any further conceptual analysis of subjective probability, we would like to offer a rough characterization of the core of what the different uses of subjective probability have in common, across these disciplines that invoke the notion:

(A) The event space $\mathcal{E}$ and the probability values should capture an agent's subjective "viewpoint" on the world; that is, how possibilities "look" from the perspective of the agent, with regard to their (relative) plausibility.

(B) The probability values should be in some way responsive to data or evidence.

(C) Probabilities should combine with some notion of *utility* or *desirability* to guide the agent's actions and choices, and perhaps more generally play a role in the agent's practical reasoning (viz. planning, etc.).

Beyond this common core, a number of authors have offered further *analyses* and *definitions* of subjective probability.[1] Many attempt to reduce the concept to some more primitive or easily understood (or measured) phenomenon, and in particular to some kind of behavior, or behavioral disposition. We can identify three important and influential analyses of this sort. First, a very common and prominent proposal has been to establish conditions on preference relations—e.g., over gambles, or relative likelihood statements—such that any agent with preferences satisfying such-and-such conditions can be *represented* as harboring certain subjective probability judgments (Ramsey 1931, Savage 1954, Anscombe and Aumann 1963 are classic examples). The agent's probabilities can then be identified with one of these representing probability functions. A second proposal is to identify an agent's probabilities with their *betting quotients*: roughly, the odds on bets involving elements of $\mathcal{E}$ that the agent deems fair (de Finetti, 1974). This construal plays naturally into the so called Dutch book argument for *justifying* the standard probability axioms for subjective probability. A third view, defended in different ways by both Lewis (1974) and Davidson (1975), claims subjective probabilities form part of a rational reconstruction by an *interpreter*, of what guides an agent's overall behavior. The interpreter observes the agent's actions and choices, and ascribes probabilities and utilities that most effectively *rationalize* these actions and choices, under the assumption that she is maximizing her expected utility.

All of these proposals make good on (A)-(C), but go beyond this conceptual core in saying exactly what theoretical role subjective probability is meant to play, including how it is to be measured or determined. While these proposals may work for characterizing scenarios like 1 and 2 above,[2] they are arguably less promising as characterizations of the relevant states in examples like 4, 5 and 6. Monkey brains do not literally assess fairness of bets, and the visual system does not express preferences over a rich set of gambles. Moreover, even if we could somehow apply these frameworks to the mechanisms in question, it is not obvious that the probabilities thus obtained would coincide with what in fact explains how these mechanisms work, that is, how in fact the brain uses probabilistic representations to produce the behaviors in question, if it does. The proposals mentioned above all have the advantage that they can be applied even in cases where probabilities are nowhere explicitly represented in the agent's mind. Subjective probabilities can be constructed and ascribed in any case where an agent exhibits *behavior* of the requisite kind, quite independent of what is going on in the agent's mind. However, this flexibility is a potential downside when we suspect that perhaps the brain does incorporate probabilistic representation—we then want to understand these representations as such.

What is a useful concept of subjective probability for the purpose of cognitive science and psychology, where we are interested in how, concretely, the human mind copes with uncertainty? We claim that further analysis of the *concept* itself, beyond (A)-(C), is unnecessary. If we can find a psychological phenomenon that is properly modeled

[1] (C) is usually, but not always, made explicit in analyses of probability. As many have argued (e.g., James 1890; Ramsey 1931; Williamson 2015, etc.), a notion of idle belief, totally disconnected from any kind of choice or action, seems suspect. Some evidence in neuroscience suggests that, for low-level cognition, representation of utility and probabilistic expectation cannot be separated (Gershman and Daw, 2012). Others insist that subjective probability is *conceptually* separable from any relation it might have to action or choice (Eriksson and Hájek, 2007). Though we assume (C) as part of the concept, not much will hinge on whether probability and choice are really conceptually separable.

[2] Of course, the many problems with these proposals, even for the types of scenarios for which they are designed, are well known. See Zynda (2000) on representation theorems, and Eriksson and Hájek (2007) for many of the central puzzles and problems with all conceptual analyses considered in the literature.

with the probability calculus, and which plays an appropriate role in the agent's psychology viz. (A)-(C), then we will have identified a useful and substantive species of subjective probability, one apt to play an important part in psychological theorizing about, and investigation of, human inference, problem solving, decision making, and other aspects of cognition. We will not claim that this is the only useful species of subjective probability, even within psychology—for instance, our proposal will not directly apply in scenarios like 1 and 2, where one still might like to invoke subjective probability—but that it has the potential to be an important and foundational one—indeed centrally operative in scenarios like 4 and 6, and in ordinary deliberative scenarios like 3 as well.

We will therefore not be analyzing the concept of subjective probability in terms of other concepts. However, we will be *identifying* subjective probabilities in the relevant sense with independent, "objective" propensities. As mentioned above, the propensity interpretation of probability is sometimes offered as an alternative to subjective interpretations. But we will not be analyzing this notion of propensity any further either. It would be consistent with our view, for example, to analyze probabilistic propensities in terms of some other notion of subjective probability (as is done, e.g., in de Finetti 1974). All we will require is that we can somehow make sense of such (relatively objective, or at least widely intersubjective) statements as, "The roulette wheel has a $1/38$ chance of landing on 7." This will become clearer in what follows.

3 Subjective versus Psychological Probability

The very idea that some psychologically primitive structures could play the role of real subjective probabilities—consistent, coherent probabilistic representations in the usual sense—has been taken by many as a non-starter. Early probabilists such as Ramsey and de Finetti did not venture to speculate about what happens inside the mind, leading to their behavioristic analyses. Since then, much of the best known psychological work that has been done on judgment and choice under uncertainty shows subjects systematically behaving at odds with the probability calculus, leading many to doubt that there is any genuine representation of probability to be uncovered in the mind.

An early but representative view on the matter is that expressed by I. J. Good (see Good 1983), who distinguished subjective probability from *psychological* probability. Psychological probability is "the kind of probability that can be inferred to some extent from your behavior, including your verbal communications," while subjective probability is "psychological probability modified by the attempt to achieve consistency, when a theory of probability is used combined with mature judgment" (73-74). According to Good, psychological probabilities would not be expected to satisfy the probability axioms. Still, these "untutored intuitions" may provide a basis on which to build up a system of genuine probabilities encoding the person's "considered judgments," which will not only be consistent, but will reflect any further thoughts that come up in the course of explicit reflection and deliberation. As he put it,

> Judgment and logic must be combined, because although the human brain is clever at perceiving facts, it is also clever in the rationalization of falsity for the sake of its equilibrium. You *can* make bad judgments so you need a black box to check your subjectivism and to make it more objective. (Good, 1983, 26)

For instance, in scenario 2 from the previous section, imagine that the patient's primary symptom is a soar throat, and the doctor realizes that this could be caused by strep or a virus.[3] Penicillin treats strep, but also leads to several days of discomfort, in extreme cases may cause death, and in general encourages penicillin-resistant bacteria. Prescribing rest would be appropriate if the patient has a virus, but if it is strep, this could lead to serious conditions such as rheumatic heart disease. In such a scenario the doctor may well have an initial "gut feeling" about how probable all of these outcomes are, and therefore what the right treatment might be. But given both the severity and the complexity of the situation, it would seem appropriate to step back and formulate the problem in a more explicit way, making calculations where needed.[4] The doctor may look at statistical data, including specific data on patients with similar characteristics, perhaps take some tests and make calculations using Bayes Rule, and somehow combine all of this to come up with probability values for each possible outcome that summarize all of the evidence as far as possible.

The resulting probabilities would be subjective probabilities in Good's sense (and, as they presumably satisfy (A)-(C), also in our laxer sense), but such cases are outside the purview of our main subject matter in this paper. Contrast

[3] This is an example from Raiffa (1968).

[4] There are, of course, those who argue that intuition outstrips explicit calculation, particularly in cases like these (Dreyfus and Dreyfus, 1986).

such examples with judgments like the following:[5] observing that a person has lived to age 10, what would you predict that person's lifespan to be? What about a person you observe at age 75? Or at age 92? As in the previous case, one could set up an explicit probability model, recording actual data and making a calculation. However, even without doing this, people turn out to be surprisingly adept at such prediction problems, showing sensitivity to subtle differences in distribution forms, including in cases where the predictions turn out to be inaccurate, e.g., because of misinformation (Griffiths and Tenenbaum, 2006). In this work, and in much other recent work on probabilistic cognition (see, e.g., Griffiths et al. 2010; Tenenbaum et al. 2011 for reviews), one is reminded of Laplace's famous quotation: "The theory of probability is at bottom nothing more than good sense reduced to a calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it." These intuitive judgments—something closer to what Good referred to as judgments or psychological probabilities, but not in general explicit numerical probability estimates—is the target of interest in this paper. The question is, what kind of representation and calculation could underly such intuitive judgments, as well as other aspects of (largely unconscious, but high-level) predictive cognition?

4 Doubts about Probabilistic Representation

Before proceeding with the proposal, it is worth briefly reviewing why many have been skeptical of probabilistic mental representation. Good alluded to the brain's "rationalization of falsity," but this is only the tip of the iceberg. We might locate the primary sources of doubt around three key issues: complexity, imprecision, and coherence.

To illustrate each of these worries concretely, let us consider one more simple type of automatic, "untutored" judgment.[6] Suppose I am describing a room in someone's home. Before I say anything about the room, how likely would you say it is to contain a chair? And how likely would you say it is to contain a stove? Whether or not you would be willing to assign numerical values to these possibilities, presumably you would judge the first to be *a priori* more likely than the second. Suppose I now tell you that the room contains a window and a lamp. This might make the possibility of there being a chair even more likely, while perhaps making the possibility of a stove yet less likely. If instead I had mentioned that there is a sink in the room, that might reverse things, making the stove more likely than the chair. We could go on for some time asking such questions. As with examples like 3 and 4 and the lifespan example above, people do seem to have flexible, robust intuitive responses to these types of questions. Could this possibly be upheld by probabilistic representations?

The best known objection to the idea of probabilistic representation and calculation in the mind is that it is too complex. Given the shear number of propositions, events, and situations we can evidently entertain, many have found it implausible that the mind could have encoded numerical values corresponding to probabilities for all of them. The situation looks even worse when we consider the idea that we may need to entertain arbitrary *conditional* probabilities, for learning or inference, as in the examples above. Many recent authors in cognitive science have expressed this worry about probabilistic models and computation (e.g., Gigerenzer and Goldstein 1996; Kwisthout et al. 2008). In the philosophical literature, Gilbert Harman famously gave voice to this skepticism:

> If one is to be prepared for various possible conditionalizations, then for every proposition P one wants to update, one must already have assigned probabilities to various conjunctions of P together with one or more of the possible evidence propositions and/or their denials. Unhappily, this leads to a combinatorial explosion. (Harman, 1986, 25-26)

That is, even if we have only a relatively small number of basic propositions that are assigned probability values, if we also need to consider probabilities for these propositions conditioned on various information, that may dramatically increase the number of probabilities we need to track.

In the room example, suppose we only consider 5 possible pieces of furniture: a chair, a lamp, a window, a sink, and a stove. In this case, there are 32 possible combinations of these objects, i.e., 32 possible rooms, and 80 different conditional probabilities we could ask about, e.g., whether there is likely to be a chair given that there is a window. Even in this very simple small-scale example it may seem farfetched to suppose that we have encoded probabilities

[5] This example is taken from Griffiths and Tenenbaum (2006). Examples 3 and 4 from above could be used to illustrate the same point.

[6] We will use this scenario, a simplified version of one from Rumelhart et al. (1986), as a running example through the next few sections.

corresponding to all of these possible queries. In Rumelhart et al.'s original scenario—only slightly more realistic, with 40 possible components of a room—there are over a trillion possible rooms, and many more conceivable conditional probabilities one ought to be able to query. This is to speak only of space or storage requirements, not to mention the time required for probabilistic computation. As Harman (1986) concluded, at least for such "everyday" unconscious predictions and judgments, subjective probabilities "are and have to be implicit rather than explicit" (36). This still leaves us with the question of what that would mean.

A second criticism of probabilistic approaches to the mind concerns the idea that judgments under uncertainty can be quantified in such a precise way as to be captured by a unique real number value, as is assumed in classical probability. This criticism is also common, and was nicely dramatized by Suppes:

> Almost everyone who has thought about the problems of measuring beliefs in the tradition of subjective probability or Bayesian statistical procedures concedes some uneasiness with the problem of always asking for the next decimal of accuracy in the prior estimation of a probability. (Suppes, 1974, 160)

Worries of this nature have motivated a number of alternative formalisms, including so called *imprecise probabilities* (Good 1983, Suppes 1974, Walley 1991, *inter alia*), designed to allow uncertainties to range within an interval, for example. Others have taken such worries to cast doubt on the usefulness of numerical representations altogether.

In the room example, it may seem preposterous to ask whether someone's subjective probability for there being a chair in a room with a window and a lamp is 0.81 or 0.81001. On the one hand, it is difficult to imagine how we could elicit such precise information in a meaningful way. On the other hand, when introspecting, people simply do not feel as though they have such definite opinions about ordinary propositions of this sort.

Finally, the long line of experimental work beginning with early seminal papers by Tversky and Kahneman, e.g., (1974), has been taken by many to show definitively that people do not reason using probability because they are in fact incoherent. For instance, in the well known conjunction fallacy (Tversky and Kahneman, 1983), subjects declare propositions of the form *A and B* to be strictly more probable than *A* alone. Returning to the room example, we might ask which is more likely: that the room has a stove, or that the room has a stove and a sink? If results across many different domains are any indication, a significant number of people would affirm the latter, which contradicts the probability axioms.[7] The conjunction fallacy and the long list of related empirical findings have generated a sizable literature, and we do not intend to give a new analysis in any of what follows. However, it is important to say why results suggesting that people are evidently not perfect probabilistic reasoners are at least consistent with a view that takes probabilistic representation very seriously. To some extent the above remarks about implicit versus explicit probabilistic representation address this point, but we shall say a bit more about it below (§7, §9).

5 Representing Probabilities with Propensities

The Sampling Hypothesis is based on the notion of a probabilistic generative process. It is common in cognitive science to assume that minds are able to construct internal representations of a given domain, sometimes called *intuitive theories* or *intuitive models* (see, e.g., Gopnik et al. 2004; Tenenbaum et al. 2011; Clark 2013; Goodman et al. 2014, etc.). In many cases, the primary function of these models is to *generate instances* of some concept, event, or other in-principle-observable data associated with that domain, with different probabilities.

We can think of these generative processes as defining a random variable (or set of random variables) taking on values in some given set $\mathcal{V}$. In our toy example of the room schema, $\mathcal{V}_{\text{room}}$ might consist of the set of 32 possible room configurations, i.e., combinations of objects from among chair, lamp, window, stove, and sink. A generative model $\mathcal{M}$ for this domain would define a process that generates room instances. For instance, on one "run" of the model $\mathcal{M}$, it might generate a room with a chair, a lamp, and a window. On the next, $\mathcal{M}$ might generate a room with a stove, a sink, and a window. And so on. Each such outcome has a probability of being generated. Thus, the generative model implicitly defines probabilities for each possible room type.

Given the distribution P that $\mathcal{M}$ defines on the sample space $\mathcal{V}$, we can then speak about probabilities of arbitrary events, i.e., subsets of $\mathcal{V}$. For example, $P(\mathsf{chair})$, the probability of there being a chair in a randomly generated room,

[7] This is particularly bad news for anyone who wants to define subjective probability on the basis of representation theorems. There is obviously no probability function that will agree with such an ordering, since $P(A \,\&\, B) \leq P(A)$ for any A and B.

is given by the marginal probability obtained by summing over the probabilities of all those room instances with chairs. Conditional probabilities such as $P(\mathsf{stove} \mid \mathsf{sink})$—the probability of there being a stove provided there is a sink, defined in the standard way: $P(A \mid B) = P(A \,\&\, B)/P(B)$—would not be given by any calculation using conjunction explicitly, but rather by the probability of generating a room with a stove, provided we only generate rooms with sinks.

The hypothesis, simply put, is that these *propensities* of a model $\mathcal{M}$ play the role of subjective probabilities. The mind encodes probabilistic information directly by its ability to support generative processes whose probabilistic dynamics implicitly define the distribution. This hypothesis, in one form or another, has been suggested recently in neuroscience (Fiser et al., 2010; Berkes et al., 2011), cognitive psychology (Shi et al., 2010; Vul, 2010), and developmental psychology (Denison et al., 2013); representatives of this and other work will be discussed in what follows. This more recent work builds on several different traditions within cognitive science, including the theory of stochastic choice (Thurstone, 1927; Luce, 1959; Luce and Suppes, 1965; McFadden, 1973), mental models theory (Craik, 1943; Johnson-Laird, 1983), and work in neural networks and in machine learning (Rumelhart et al., 1986; MacKay, 2003; Koller and Friedman, 2009).

What would a physical process implicitly representing probabilities in this way look like? To give the basic intuition, an early example of a sampling machine was given by Sir Frances Galton, who used the *quincunx machine*, later called the Galton box (Fig. 1), to illustrate various probabilistic concepts.[8]

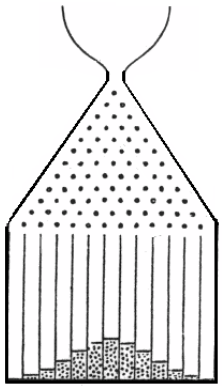

Figure 1: The Galton box (from Galton 1889)

The probability that a pebble will drop in a given slot is roughly proportional to the probability associated to that interval in a normal distribution. After enough pebbles have been dropped, with high probability the distribution resembles the normal distribution. In this picture, we can think of the different slots as the possible hypotheses—the sample space $\mathcal{V}$—and each pebble as a sample from the underlying (approximately normal) distribution over $\mathcal{V}$.

Of course, this is meant only as an illustration. For a more serious (if still toy) example, let us return to the room scenario. We can model such judgments using a *Boltzmann Machine* (Rumelhart et al., 1986), which is a simple neural network $\mathcal{N} = \langle N, \mathbf{W} \rangle$ given by:

- a set N of binary nodes, in this case $N_{\text{room}} = \{\mathsf{chair}, \mathsf{lamp}, \mathsf{window}, \mathsf{stove}, \mathsf{sink}\}$;
- a symmetric weight function $\mathbf{W} : N \times N \to \mathbb{R}$, s.t. $W_{i,i} = 0$ and $W_{i,j} = W_{j,i}$ for all $i, j \in N$.

Plausible parameters for the particular example we have been discussing are depicted in Figure 2. Intuitively, the weights between nodes represent the (positive or negative) correlation between what the nodes represent. Thus, stoves and chairs are anti-correlated, while stoves and sinks are correlated. The basic *activation function* for the Boltzmann machine is a stochastic update rule that determines whether a node i will be on or off, as a function of whether other nodes are currently on or off, and the weights between those nodes and i. In particular, a node i is randomly chosen,

[8] Cf. Vul (2010), who earlier used this illustration.

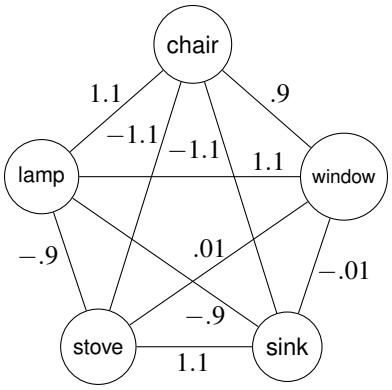

Figure 2: Boltzmann Machine $\mathcal{N}_{\text{room}}$ for the simple room schema

and is turned on (or remains on) with probability given by the familiar logistic function:

$$\frac{1}{1+e^{-net_i}}$$

where $net_i = \sum_j W_{i,j}\mathbb{I}_j$, and $\mathbb{I}_j$ is an indicator function, equal to 1 if x_j is currently activated, 0 otherwise.

As explained in Appendix A, this simple activation rule can be seen as carrying out the so called *Gibbs Sampling* algorithm on an underlying distribution given by an associated energy function. This *Boltzmann distribution* gives us a well defined probability P_{room} on sample space $\mathcal{V}_{\text{room}}$, which we can use to model the judgments discussed above. With the weights as in Fig. 2 we can calculate that

$$P_{\text{room}}(\mathsf{chair} \mid \mathsf{window},\ \mathsf{lamp}) = 0.81 \quad \text{while} \quad P_{\text{room}}(\mathsf{chair} \mid \mathsf{stove},\ \mathsf{sink}) = 0.24.$$

Furthermore, while the prior probability of there being a stove is low (0.30), provided there is a sink the probability of there being a stove as well is above chance (0.56).

The details of this particular example, including the particular parameters, are not important.[9] There are two main ideas we want to convey with it. The first is that it shows one way a relatively complex distribution, including a great deal of information about *conditional* probabilities, can be encoded with a very compact representation. The network depicted in Figure 2 is quite small compared to the number of (conditional) probabilities it implicitly defines.

Second, and most importantly, $\mathcal{N}_{\text{room}}$ represents the distribution P_{room} precisely in the sense that the probability of this machine outputting a given value $v \in \mathcal{V}_{\text{room}}$ is $P_{\text{room}}(v)$. That is, if we apply the activation function as described above for a sufficient number of steps, the resulting vector of values will amount to a *sample* from distribution P_{room}. Likewise, if we do this with some of the nodes "clamped" to remain on, then the network will sample from the corresponding conditional distribution. E.g., clamping window and lamp, we can use the machine to sample from $P_{\text{room}}(\mathsf{chair} \mid \mathsf{window},\ \mathsf{lamp})$. Indeed, given how Boltzmann Machines and related tools have been used in cognitive and neuroscientific modeling (see, e.g., Churchland and Sejnowski 1994), we would expect sampling in this way to be the primary means of extracting probabilistic information from the underlying distribution. The exact calculations above are quite involved and are best done by computer. They do not describe information that is itself readily available to some other mechanism which might use this machine. Instead, these probabilities describe information that is implicit in the network, and which can be approximated by running the network in the manner described.

To a certain extent, these observations already show how we meet the complexity challenge raised by Harman and others. It is by making probabilities implicitly, rather than explicitly, represented, just as Harman suggested.[10] The

[9] Presumably, a more realistic probabilistic model of this domain would have to include higher-order correlations as well. For instance, if we added minibar to the network, then while sink and chair are anti-correlated in general, in the presence of minibar they might be correlated. This kind of structure can be captured by more general Markov random fields. See next section.

[10] Though this is a very different proposal from what Harman suggested (Harman, 1986, Ch. 1). See Millgram (1991) for a response to Harman also drawing on (approximate inference for) graphical models. Millgram goes further, questioning whether any "hardness argument" could ever cast doubt on the idea that we *ought* to reason in a way consistent with numerical probability.

network $\mathcal{N}_{\text{room}}$ defines all 80 conditional probabilities associated with this scenario in an extremely compact format. In fact, there are formal results to this effect, showing that the *efficiently samplable* class of distributions, under a suitable formalization, are a proper superset of the class of *efficiently computable* distributions, under standard complexity assumptions (e.g., P $\neq$ NP; see Arora and Barak 2009, Ch. 15). For many distributions of interest, performing even "near perfect" calculations by sampling is still intractable (Kwisthout et al., 2008), so we should expect performance to be far from perfect in general. The hypothesis is that the probabilities in question are nonetheless to be understood as *really there*, represented implicitly, to be potentially used in making predictions and decisions.

6 Some Clarifications

What is intended to be the scope of the Sampling Hypothesis? In particular, what types of distributions should we assume are (or could be) represented in this way, what is the mode or medium of this representation, and how might these implicitly defined probabilities in fact be used by an agent?

The Boltzmann Machine described above can only represent certain kinds of distributions. In the parlance of graphical models, it is a specific kind of *Markov random field*.[11] These are essentially undirected models in the sense that all probabilistic dependencies are assumed to be symmetric. There are many other graphical models that have been used in cognitive science and artificial intelligence (Pearl, 1988; Koller and Friedman, 2009), of which the best known are Bayesian networks. All of these models have in common the idea of reducing the amount of information from the full probability table that needs to be encoded, taking advantage of (or just assuming) independencies between variables. They do this in different ways. Bayes nets, for example, require *asymmetric* dependencies, and unlike Markov random fields may not contain cycles. Some recent authors have argued that in cognitive science we need significantly more powerful probabilistic representations, e.g., arising from full (probabilistic) programming languages, with the possibility of probabilistic recursion (Freer et al., 2012; Goodman et al., 2014) and more powerful kinds of abstraction (Chater and Manning, 2006; Tenenbaum et al., 2011). The distinctions among these models are not important for our purposes, and we could just as well have used any graphical model with an appropriate sampling algorithm to illustrate the main points.

Of course, to the extent that a formalism can be implemented in a neurally inspired physical mechanism, that speaks in its favor. In this sense, the Boltzmann Machine is a suggestive example, but there are others (see §8).

Independent of the abstract computational form of these generative processes, there is also an obvious question of how the Sampling Hypothesis relates to debates about where and how computations are performed in the brain. There is compelling evidence that the mind uses episodic memory traces to simulate possible future events for the purpose of prediction (Schacter et al. 2008, who suggest this may even be the primary function of episodic memory).[12] This is certainly compatible with the Sampling Hypothesis. It should also be clear that the hypothesis is compatible with views according to which reasoning and prediction are grounded in perceptual/motor processes (Barsalou 1999). Merely introspecting on one's inferences about the room schema, for example, seems to reveal (at least accompanying) visual mental imagery of possible rooms, and perhaps other sensory modalities. Connections between computational models and neural implementation for low-level cognition will be discussed briefly in §8, but it is important to stress that the hypothesis is not anchored to any very specific proposal.

Next, we have not yet said anything precise about how these representations of uncertainty are used by a subject. Here is a decision rule illustrating how samples could be used for a simple elicitation task (cf. Vul et al. 2014):[13]

> DECISION RULE A: Given a task to report which of n hypotheses $\mathcal{V} = \{H_1, \ldots, H_n\}$ is true, and a generative model $\mathcal{M}$ with $\mathcal{V}$ as possible return values, take R samples from $\mathcal{M}$ and let BEST be the set of hypotheses that receive the largest number of samples. Return any $H \in$ BEST with probability $\frac{1}{|\text{BEST}|}$.

There will obviously be such a rule for any number R of samples. For instance, using the Boltzmann Machine in Fig. 2 to answer the question, "Will a room with a stove also have a sink?", DECISION RULE A with $R = 3$ would have

[11] Even within the class of Markov random fields, Boltzmann Machines are a small subclass, representing only distributions that can be written with pairwise potential functions.

[12] See also Stewart et al. (2006) on an application of this idea to decision making.

[13] This only works for relatively small discrete hypothesis spaces. For continuous spaces a natural alternative proposal would be to construct a density estimate and return the mean, for instance.

us clamp the stove node and run the network for a while before seeing whether the sink node is actived, repeating this three times.[14] With very high probability, at least two of those three runs would have sink actived, so DECISION RULE A would likely return a positive response.

DECISION RULE A makes sense for problems with 0/1 utility functions, e.g., when one is only interested in obtaining a "correct" response, in which case expected utility and probability coincide. In more general situations with varying cardinal utility, it is possible to generalize DECISION RULE A to the following DECISION RULE B:

> DECISION RULE B: Suppose we are given a task with possible states $\mathcal{V} = \{H_1, \ldots, H_n\}$ and a generative model $\mathcal{M}$ with $\mathcal{V}$ as possible return values. We further assume a set of actions $\mathcal{A} = \{A_1, \ldots, A_m\}$ and a utility function $u : \mathcal{A} \times \mathcal{V} \to \mathbb{R}$. To select an action, take R samples, $H^{(1)}$, $\ldots$, $H^{(R)}$, using $\mathcal{M}$, and let BEST be the set of actions that receive the largest summed utilities, i.e.,
>
> $$\text{BEST} = \{A_j : \sum_{i=1}^{R} u(A_j, H^{(i)}) \text{ is maximal}\}.$$
>
> Take action $A_j \in \text{BEST}$ with probability $\frac{1}{|\text{BEST}|}$.

Following DECISION RULE B, a number of sample states are drawn, the utilities for all of the candidate actions are summed over these sample states, and an action is chosen with the highest summed utility. It is easy to see that as R increases, the rule more closely approaches a perfect expected utility calculation. DECISION RULE B is of course just one possible extension of DECISION RULE A to this setting.[15]

It is important that the propensities associated with the hypothesized mechanisms can reasonably be interpreted as genuine subjective probabilities, and that the outputs of these mechanisms can therefore be seen as genuine samples from a meaningful distribution. No one doubts the mind's ability to come up with instances of a concept, to imagine possible future events, and so on. It is moreover obvious that on any given occasion we could associate a probability distribution with the possible outcomes of these processes. The substantive claim is that these sampling propensities can sometimes be understood as encoding the subject's own uncertainty. This will be founded on some combination of how these processes relate to the agent's perceptions and observations, and especially how they are used in producing verbal and other behavior, presumably using a decision rule similar to those given above. In other words, we will need to show that (A), (B), and (C) in our characterization of subjective probability are all met. If use of DECISION RULE A or B is observed in subjects' behavior, this will be good evidence that (A) and (C) are met.

How wide-ranging is the Sampling Hypothesis supposed to be? Are we to suppose that every judgment under uncertainty is the result of a sampling process using some internal generative model, invoking something like DECISION RULE A or B? Such an assumption would obviously be too strong. It is easy to come up with examples where a person might have an intuitive plausibility judgment without any capability, consciously or unconsciously, of considering alternative scenarios. For instance, I might find it very likely that the computational complexity classes P and NP are distinct. But this is not because I can "imagine" more scenarios in which they are distinct than in which they are equal. I am rather moved by expert opinion, mathematical evidence supporting the claim, and so on. How judgments of this sort work in general is well beyond the scope of this paper (though we will consider further such examples in §11). We take it as evident that judgments like this are not grounded in sampling mechanisms of the sort under discussion.

It is a difficult challenge to characterize the types of inferences for which we would predict sampling to be operative (cf. §11). The full extent of sampling in human cognition is of course an empirical question. It may therefore be more illuminating to let the empirical evidence speak for itself. First, however, it will be useful to make one final clarification concerning the relation of the Sampling Hypothesis to the so called Bayesian program in psychology.

7 Sampling and Bayes

The Sampling Hypothesis has been most explicitly explored in the literature on Bayesian psychology (e.g., Vul 2010; Denison et al. 2013; Griffiths et al. 2015, *inter alia*). There is good reason for this. Probabilistic models of psycholog-

[14] For the Boltzmann Machine, which defines a *biased* sampler (see §7, §9, Appendix A), one must specify how many iterations to run before the current state is returned as a single, bona fide sample.

[15] It is also conceivable that the actions themselves could be incorporated into the generative process. See, e.g., Solway and Botvinick (2012) and the idea of *planning as inference*, where an action is inferred from a model by conditioning on the action having desirable consequences.

ical phenomena are often quite complex, and computing with conditional distributions can often be intractable. For instance, in using Bayes Rule,

$$P(H|E) = \frac{1}{Z} P(E|H)P(H) ,$$

computing the normalizing constant $Z = \sum_{H' \in \mathcal{V}} P(E|H')P(H')$ can be very costly if the number of possible hypotheses in $\mathcal{V}$ is large. In place of exact calculation, it is sometimes possible to approximate probabilistic computations using samples from the distribution, which may be relatively easy to produce. Suppose, for instance, one needs to calculate an expectation Φ of a function $\phi(X)$ under distribution $P(X)$,

$$\Phi = \int P(X)\phi(X)\, dX .$$

For instance, in an expected utility calculation for an action A, we would have $\phi(X) = u(A,X)$. Provided one can effectively generate samples $\{X^{(r)}\}_{r=1}^{R}$ from $P(X)$, the expectation Φ can be approximated by

$$\hat{\Phi} = \frac{1}{R} \sum_{r} \phi(X^{(r)}) .$$

The Law of Large Numbers guarantees $\hat{\Phi}$ will converge to Φ as R goes to infinity. Moreover, the variance of $\hat{\Phi}$ decreases as R increases, which means the more samples taken, the closer the estimate is expected to be to the target distribution. Recall the Galton box (Fig. 1), where this can be easily visualized. In the case of expected utility, this is just another description of the computation performed using DECISION RULE B.

Computational applications of probability have led to efficient algorithms for approximate sampling, collectively called *Monte Carlo methods*, of which Gibbs sampling is a prime example (MacKay, 2003, Ch. 29). Since sampling exactly from a distribution is typically as hard as exact calculation, these methods produce biased samples, in which the sequence of random variables is correlated. Under broad conditions it can be shown that this bias washes out after enough iterations of the procedure. At the same time, these biases have themselves been used to explain certain aspects of behavior, as we will discuss below in §8 and §9.

While a connectionist modeler might use a Boltzmann Machine and note incidentally that it embodies an underlying probability distribution over states, the Bayesian modeler usually begins with the probability distribution before exploring more concrete algorithms (Anderson, 1990; Griffiths et al., 2010, 2015). Several recent authors in this literature have proposed that Monte Carlo methods and related stochastic sampling algorithms provide *algorithmic level* hypotheses corresponding to the *computational level* models at which ideal Bayesian analyses are targeted, in Marr's sense of these terms.[16] They allow people to approximate Bayesian inferences and decisions using a large enough number of samples as proxy for explicit calculations. Indeed, some of the most compelling experimental demonstrations of the viability of the Sampling Hypothesis stem from these proposals (§8-§10), which additionally purport to show the distributions from which people are sampling are rational, e.g., by virtue of being appropriately conditioned on data.[17]

What is the relation between the Bayesian program and the Sampling Hypothesis? Let us take the Bayesian program (the descriptive, not prescriptive, version) to consist in two claims:

(1) Mental states can be described in terms of (in some sense coherent) probability distributions;

(2) Learning and updating amount to conditioning an appropriate distribution.

[16] See Marr (1982), and also Anderson (1990). Compare this with the following quotation from Churchland and Sejnowski (1994):

> In the Boltzmann Machine, matters of computation, of algorithm, and of implementation are not readily separable. It is the very physical configuration of the input that directly encodes the computational problem, and the algorithm is nothing other than the very process whereby the physical system settles into the solution. (92)

For the Bayesian psychologist, the mere specification of the machine leaves out a crucial level of computational *explanation*: understanding what function the machine is computing, or "attempting" to compute (Griffiths et al., 2010).

[17] Some have pointed out that many Bayesian models in cognitive science (provably) cannot be tractably approximated, which has been taken to throw the rationality claim into question (Kwisthout et al., 2008). The general trend has been to back off to a notion of *bounded* rationality, though there remain difficult and interesting questions on this front (Griffiths et al., 2015; Icard, 2014).

The Sampling Hypothesis is committed to some version of (1), but is less closely tied to (2).

Concerning (1), in a sense coherence is built right into the Sampling Hypothesis with subjective probabilities being identified with physical propensities. A generative model $\mathcal{M}$ implicitly defines a distribution on a sample space $\mathcal{V}$. We can extend this to a coherent distribution over a full event space as we did in §5 when talking about conditional probabilities of complex events in the room model. Such a model characterizes knowledge implicit in a subject's mind, though by itself it tells us nothing about how a subject will use the model. For that we need something like DECISION RULE A. Given that sampling is inherently noisy, we might expect a subject using such a rule to give inconsistent responses to the same question on different occasions, even when knowledge of the domain has not changed. This is indeed what we find (see §10). Thus, characterizing a concrete psychological mechanism using a coherent probability distribution certainly does not rule out the possibility that we would want to characterize the agent's *behavior* as incoherent or inconsistent.[18] Additional sources of potential incoherence or inconsistency stem from the possibility that a subject will construct two different models to answer the same question, e.g., on account of framing effects (Tversky and Kahneman, 1974), or more generally will construct the "wrong" internal model, or use the right model in the "wrong" way. One might speculate that something in this direction underlies the conjunction fallacy (cf. §9).

As for (2), suppose we have established that an agent's prior probability in some situation is given by a generative model corresponding to distribution $P(X)$. Then upon receiving some information E, (2) says that the agent's new model of the situation should correspond to distribution $P(X|E)$. In the Boltzmann Machine this required only a simple adjustment to the network, namely clamping the nodes corresponding to E. And to repeat, much psychological work suggests that in many cases subjects do condition according to Bayes Rule. At the same time, few would venture that people are ideal Bayesian learners, with their current state of knowledge the result of a lifetime of perfect updates to an initial (massive) prior probability distribution. The use of some Monte Carlo methods, such as the particle filter (see §9 below), assume that subjects will settle on *locally approximately optimal* hypotheses, and use those as assumptions in future inferences rather than always reconsidering whether those earlier hypotheses are still supported by the data, to take just one example. Other authors have also stressed the local nature of prediction and judgment, both in the sense of *anatomical* or *spatial* locality (Kruschke, 2006), as well as locality in *subject matter* (Gaifman, 2004; Icard and Goodman, 2015), resulting in the prevalent flouting of Carnap's (1947) Principle of Total Evidence, according to which judgments should be made taking into account *all* of the evidence available. The full extent to which people are Bayesian in the sense of (2) is in large part an empirical question. The Sampling Hypothesis by itself is consistent with a range of possible answers to this question, as long as we can construe the probabilities as being responsive to data in some manner and to some reasonable degree (in line with (B) from §2).

8 The Neural Basis of Subjective Probability

If the Sampling Hypothesis is to be taken seriously as telling us something about how probabilities are represented in the mind, it ought to be at least consistent with what we know about neural representation. The neural basis of subjective probability is a matter of some controversy in contemporary neuroscience (Knill and Pouget, 2004; Vilares and Kording, 2011). It is generally accepted that neural firings are inherently noisy, and one of the central questions is how this noise might be used to the brain's advantage. Noise is certainly suggestive of something like sampling, but there are several alternative proposals. Perhaps the most prominent conjecture is the *coding hypothesis*, according to which spike rates of neurons directly and explicitly represent probabilistic numerical quantities, such as likelihood ratios or parameters of density functions. Firing rates have been shown to represent other continuous quantities such as direction of movement, and probability is assumed to be simply another type of quantity that can be represented in a similar way. Sophisticated proposals show how an appropriate kind of noise conjectured to characterize neural firings can aid in decoding (see Knill and Pouget 2004 for a review).

One version of the Coding Hypothesis has been partially corroborated by Yang and Shadlen (2007), who demonstrated that the firing rates of certain neurons in the monkey cortex can be adequately modeled by (a linear function

[18] Cf. Chater et al. (2006), Box 2, where a similar point is made.

of) the logarithm of the posterior odds ratio between two hypotheses H_1 and H_2 given some evidence E:[19]

$$\log \frac{P(H_1|E)}{P(H_2|E)}$$

The monkeys are presented with a series of shapes, each of which is associated with a randomly pre-assigned weight. At the end of a trial, the monkey saccades left or right and is rewarded with probability proportional to the summed weights for the hypothesis chosen. Strikingly, following the training phase, the monkeys' neural firing rates are successively updated after each new shape is presented, accurately reflecting the posterior probabilities, which in turn predict the monkeys' responses.

It is worth pointing out that, formally speaking, the neurons in this study could easily fit into a larger Boltzmann-like network. Recall that in the Boltzmann Machine the probability of an individual neuron x_i firing, given the current state of the other neurons x_j, is the result of applying the logistic function to the net input of neuron i:

$$\frac{1}{1+e^{-net_i}}.$$

It so happens that under the Boltzmann distribution, net_i is equal to the log odds ratio (see Appendix A). The results from Yang and Shadlen (2007) suggest that instead of applying the non-linear logistic function to net_i, the activity in these neurons is governed by a simpler, linear function of its input. This makes sense if these neurons are simply serving as summaries of computations performed elsewhere—perhaps by sampling—to be used directly for action selection. After all, as the authors mention, it is an open question where and how these computations occur. They were only able to measure the results of the computations. To that extent, it could very well be that there are roles for both coding and sampling in how the brain manages uncertainty.

Indeed, several other authors have pointed out ways coding and sampling might work in concert. As Lochmann and Deneve (2011) point out, in cases with only two possible states, as in the Yang and Shadlen experiment, spike rates can themselves be interpreted as samples from a distribution. Working in the coding framework, Moreno-Bote et al. (2011) present a sophisticated account of how the brain could sample directly from distributions represented by *population codes*, one of the most mathematically well-developed versions of the coding hypothesis. All of this shows that sampling is at least consistent with what we know about neural representation. Is there any reason to suppose that the brain actually employs such algorithms?

Most of the work on neural coding of probability focuses on low-level perceptual tasks, where the relevant distributions are relatively low-dimensional and can be described in terms of parametric models, e.g., using normal distributions. In these cases it is possible to summarize the relevant statistics by only a few values, e.g., mean and variance. However, as many have noted (Fiser et al., 2010; Vul, 2010), coding models do not easily scale up to more complex tasks such as categorization, language understanding, causal inference, or even higher-level visual processes (Lee and Mumford, 2003). For models complex enough to handle tasks like these, sampling is a favored engineering solution (MacKay, 2003). It is tempting to postulate that the brain may have happened upon similar solutions. In fact, there is suggestive behavioral and modeling evidence that this may well be the case.

One of the most intriguing empirical arguments for sampling in perception comes from *multistability phenomena*, as with ambiguous visual images where subjects report flipping back and forth between interpretations. An example of multistability is binocular rivalry, where one eye is presented with one image and the other eye with another (see Figure 3). In these cases subjects report percepts flipping back and forth, occasionally resting briefly in intermediate "hybrid" states. A quintessential feature of the Boltzmann Machine with Gibbs sampling (and many related models) is that, even after it has settled into a steady state, with positive probability it will transition to another state. If there are several energy minima in the state space (states with high probability relative to "neighboring" states), such transitions may occur often. A number of researchers have noticed that this suggests a possible analysis of multistability phenomena. Indeed, ever since von Helmholtz, it has been common to view perception as the brain's solution to an inference problem, namely guessing the latent state of the world from noisy perceptual data. To the extent that different latent states have different probabilities given the perceptual input, the Sampling Hypothesis predicts that the perceived image will amount to a sample from this distribution on interpretations. One of the attractive aspects of the multistability

[19]The results in the paper are stated in terms of the log likelihood ratio, also known as the *weight of evidence* (Good, 1950); but since the two hypotheses have equal prior probability in these experiments, these values are equal by Bayes' Theorem.

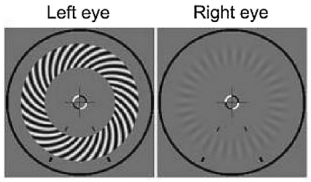

Figure 3: Binocular rivalry experimental stimuli from Gershman et al. (2012)

phenomena is that we can observe the behavior of the visual system over an extended period with the same input, as though we are witnessing multiple samples from a single underlying distribution.

In one recent paper, Gershman et al. (2012) define a probabilistic model for inferring luminance profiles from two separate retinal images—in fact based on a variation of the Boltzmann Machine—and show that the Gibbs sampling algorithm predicts a number of central empirical results from the literature on binocular rivalry and multistability.[20] For instance, they are able to explain the gamma-like distribution describing the time between switches, which had required *ad hoc* postulations in previous accounts of multistability. On their model, this is a natural consequence of the sampling dynamics and the network topology, as are several other phenomena associated with multistability.

Sampling has also been implicated in "mid-level" cognitive phenomena involving logically complex spaces, such as categorization, language understanding, causal cognition, and many others, areas where alternative computational models that work in low-level vision seem ill-suited. To give a concrete example, consider the problem of parsing a sentence. It is common to assume that sentences have some underlying ("deep") structure and that the task is to infer the appropriate structure from the surface form of the sentence. In general, there are many possible structures corresponding to any given surface form, making parsing a hard problem.[21] Levy et al. (2009) show that a model based on the *particle filter*—a Monte Carlo algorithm designed for sequential addition of data—is able to predict some of the central findings from the literature.[22] The basic idea is that a number of "particles" are maintained with different weights, corresponding to hypotheses about the correct interpretation of the sentence. As each word is perceived, a new set of (weighted) particles is sampled, probabilistically depending on the new word and the particles (and their weights) from the previous step. Among other predictions, the particle filter algorithm accounts for "garden path" sentences, like that in $(*)$ below, which subjects routinely take a long time to process.

$(*)$ The woman brought the sandwich from the kitchen tripped.

If the number of particles is small, interpretations that seem to have low probability initially—here, the interpretation on which 'the woman' is the object of 'brought'—may simply drop out, and the algorithm crashes upon processing 'tripped', requiring the analysis to start over. With the number of particles intuitively standing in for working memory, Levy et al. (2009) were able to explain much existing data on these phenomena.

9 Heuristics and Biases

So far we have reviewed evidence that sampling may be implicated in relatively automatic low- and mid-level psychological processes like vision and parsing. One might wonder to what extent high-level judgments and predictions are the product of a sampling process. That people do not always behave like ideal probability theorists in this kind of task is by now commonplace. The large body of work by Tversky and Kahneman (1974), and much work following, has identified a number of general *heuristics* that people seem to use and the resulting cognitive *biases*, and their behavioral effects are by now well established. This program has sometimes been characterized as in opposition to

[20] They build on a large literature on this topic, including earlier probabilistic and sampling-based analyses of multistability, especially Sundareswara and Schrater (2007). See Gershman et al. (2012) for other references. See also Moreno-Bote et al. (2011), cited above, for a different sampling analysis of multistability based on population coding, and Buesing et al. (2011) for extensions of Gershman et al.'s work, in a more neurally plausible setting, accounting for refractory periods.

[21] Disambiguating sentence structure is akin to disambiguation of word sense, as in one of our initial examples (4) from §2.

[22] Notably, particle filters have also been proposed as neurally plausible models of high-level vision. See, e.g., Lee and Mumford (2003).

probabilistic models, and especially "rational" probabilistic models (e.g., Gigerenzer and Goldstein 1996, *inter alia*). However, an interesting twist explored in very recent work is the idea that some of these heuristics and biases can be seen as natural consequences of sampling from appropriate distributions.

One of the classic examples introduced in Tversky and Kahneman (1974) is the *anchoring and adjustment heuristic*. In one of their experiments, they asked subjects to estimate some unknown quantity n—for instance, the percentage of African countries in the United Nations—after first rolling a roulette wheel with numbers 1-100 and having subjects say whether n is greater or less than the number on the wheel. The finding, which has been reproduced many times and in many different contexts, is that responses are highly skewed toward the number on the roulette wheel, suggesting that people take this number as an initial hypothesis ("anchor") and "adjust" it to what might seem a more reasonable value. For instance, on the UN question, the median response from subjects who saw a 10 on the wheel was 25, while the median response from subjects who saw 65 on the wheel was 45.

On the face of it, the anchoring and adjustment idea has a similar structure to Monte Carlo algorithms. The latter always begin in some initial state, i.e., with an initial hypothesis, and then "adjust" the hypothesis as the search space is explored along the constructed Markov chain. Recall the Boltzmann Machine. The network begins with some initial activation vector, with each node on or off, before stochastically updating nodes according to the logistic activation function. Clearly, if the network is only run for a brief amount of time, it will likely remain close in state space to the initial vector. This is a general feature of many Monte Carlo algorithms. Lieder et al. (2012) have demonstrated that much of the data on anchoring and adjustment can indeed be accurately modeled by assuming that people approximate sampling from the posterior distribution using a Monte Carlo algorithm that is initialized with a salient answer to a simpler question.

Other well known heuristics from this literature are at least consistent with sampling. Imagine, for instance, a general way of using sampling to determine which of two event-types E and F is more probable: draw samples to determine how likely E is, separately draw samples to determine how likely F is, and compare the respective ratios. Such a method has the feature that it could be used to compare very different event types. For instance, one might judge tomorrow's weather being sunny to be more probable than that a frog will leap a meter high, perhaps because one's mental simulations of tomorrow's weather turn up sunny more often than one's mental simulations of frog leaps turn up a meter high. If such a method were used to determine the relative probability of event-types that could be part of the same overall event space—where the mind could, but perhaps does not in fact, use the same sampling process for the two events E and F—this might explain conjunction-like fallacies associated with Tversky and Kahneman's (1974) *availability heuristic*, declaring E to be more probable than F even when E is "more specific" than F.

In one of Tversky and Kahneman's original experiments, subjects were asked to estimate how likely a word in English is to end in 'ing' and (separately) how likely a word is to have 'n' as the penultimate letter. They judged the former much more probable than the latter, even though the latter are a strict subset of the former (note the similarity to the conjunction fallacy). The received explanation is that it is much easier to probe one's memory for 'ing' words, and that subjects are using "availability" as a heuristic stand-in for probability.[23] If subjects were using a method like that described above—sampling '-ing' words and '-n-' words separately, and not noticing that one is a subtype of the other—the brute fact about human memory retrieval, that '-ing' words are easier to recall than '-n-' words, could perhaps account for the phenomenon. This is nothing more than a just-so possible explanation as it stands, but it shows that the sampling hypothesis is quite consistent with use of an availability heuristic.[24]

Rather than casting doubt on the Sampling Hypothesis, some of the observed heuristics in judgment and prediction behavior corroborate the idea that people deal with such tasks by sampling from internally constructed models. Understanding how such models are constructed and how samples are generated are important ongoing research questions.

[23]The literature following Tversky and Kahneman (1974) has investigated an ambiguity in the statement of the heuristic, as to whether probability judgments are based on the *number* of instances brought to mind, or the *perceived ease* with which examples are brought to mind. Experimental results by Schwarz et al. (1991) suggest it is the latter.

[24]It is also tempting to suggest that the conjunction fallacy itself can be explained in terms of such a method. Perhaps more *representative* (Tversky and Kahneman, 1983) instances of a concept or event-type are indeed sampled more often and easily, just as we know more *prototypical* instances of a concept often come to mind most easily in experimental settings (Rosch, 1975). While the method of separately sampling the two event types has the advantage of working in principle for any two event-types, it risks making systematic logical mistakes in cases like these. Needless to say, this vague claim would need to be systematized and tested. One might also worry, to the extent that this holds, whether the sampling propensities in question could still be said faithfully to represent the agent's degree of uncertainty. Worries of this sort will be addressed in §11 below.

Work such as that of Lieder et al. (2012)[25] provides hope that specific sampling algorithms may offer important insight into these questions. It moreover strengthens the suggestion that these samples are drawn from distributions that we can sensibly view as internally representing the subject's own uncertainty.

10 "The Crowd Within"

The various sampling algorithms—Gibbs sampling, particle filtering, etc.—each have their own characteristic biases and dynamics. As explained in the previous section, these dynamics can sometimes be shown to match aspects of empirical psychological data, accounting for deviation from purportedly normative behavior. What all of these algorithms have in common is that they may explain the prevalent phenomenon of *probability matching*.[26]

Suppose in some experiment subjects are presented with two alternative event types A and B on 70% and 30% of trials, respectively, and are asked to make a prediction about whether the next event will be an A or B event. Population-level probability matching occurs when roughly 70% of subjects respond with A and 30% with B. That is, instead of each subject always choosing the most frequently observed event type, the distribution of responses matches the empirical distribution associated with the stimuli. As Vul et al. (2014) and others have observed, this phenomenon extends beyond simple frequency or probability matching to, so to speak, *posterior matching*, where the distribution of responses actually matches some normative posterior distribution for the task under investigation. This has been observed across a number of domains, including categorization, word learning, causal learning, psychophysics, and others (see Vul et al. 2014 for a list of references).

Under the assumption that the probabilities implicit in a subject's generative model match the normative probabilities associated with a given task, the simple DECISION RULE A from §6 with $R = 1$ sample (and in fact also with $R = 2$) predicts probability matching exactly. The probability of responding with H is given by the probability of drawing H as a sample, which just is the probability of H. Vul et al. (2014) have even shown that, across many studies over several decades on probability matching, there is a correlation between the stakes of a given problem and the number of samples that would explain subjects' behavior if they are using something like DECISION RULE A. People can, and apparently do, strategically adjust the number of samples they draw before giving a response, depending on how much of a difference they could expect more samples, and thus presumably better accuracy, to make.

This explanation of probability matching arguably makes more sense than a popular alternative account, according to which probability matching arises from subjects using a *softmax decision rule* (sometimes called the *generalized Luce-Shepard rule*). In the specific case of estimation problems, the softmax rule predicts response H with probability

$$\frac{e^{v(H)/\beta}}{\sum_{H' \in \mathcal{V}} e^{v(H')/\beta}}$$

where $v(H) = \log P(H)$. For $\beta = 1$, this is equivalent to DECISION RULE A with $R = 1$ (or 2), i.e., probability matching. Also like DECISION RULE A, the softmax rule can model the gradient between probability matching and maximizing by varying the value of β. Just as the probability that a sample-based agent will guess the most probable hypothesis as the number of samples goes to infinity, as β goes to 0 the softmax probability goes to 1 for the most probable hypothesis.

While the two can be used almost interchangeably to explain probability matching behavior, if brains were to implement the softmax rule literally, they would need to be able to compute probabilities perfectly. When experimentalists use the softmax rule to predict responses, the noise is interpreted either as encoding our own uncertainty about utility (McFadden, 1973), or perhaps as random error in action execution. It is clear that Luce (1959) intended his rule merely as providing an adequate *description* of choice behavior (see also Luce and Suppes 1965), rather than as a proposal about mental computations. The rule has since been used frequently across a number of psychological domains because of its flexibility in fitting data. Sampling-based decision rules like DECISION RULE A may offer a principled, more mechanistic explanation of this behavior (Vul, 2010; Vul et al., 2014), at least in some cases. Often it seems as though the difficult aspect of a decision problem—where the brain may need to take a shortcut—is precisely

[25] See also Lieder et al. (2014) and Griffiths et al. (2015) for more work in this vein, including results relevant to the availability heuristic, specifically availability of "extreme" events (cf. §11 below).

[26] See Vulcan 2000 for a comprehensive discussion and review of work from the 20th century.

in estimating the state of the world, and what will happen if a given action is taken. Further observations about the relationship between the softmax rule and sampling-based decision rules, including a suggestion for distinguishing them qualitatively, can be found in Appendix B.

DECISION RULE A (like some possible interpretations of the softmax rule) attributes population-level variability to the inherently stochastic nature of individual choice. But probability matching at a population level is of course consistent with each individual subject following some deterministic procedure. This potential objection is especially pressing given that population-level probability matching behavior has been shown to evolve in simulated populations of completely deterministic agents (e.g., Seth 1999). Why in a given case should we suppose subjects have internalized roughly one and the same generative model and are each drawing only a few samples from it, when we may be able to explain the same aggregate behavior by individual variation (cf. Mozer et al. 2008.)?

If subjects' responses are drawn from a distribution associated with an internal generative model, we would ideally like to elicit multiple responses from the same individual and use those to estimate various statistics of that subject's assumed distribution. This would give the most direct behavioral evidence for the Sampling Hypothesis, and is essentially what we seem to have in the case of binocular rivalry. However, for cases of ordinary reasoning and prediction, this is in practice complicated by the fact that subjects might remember their earlier responses and not resample. Intriguingly, earlier experimental investigations of choice behavior showed that subjects would nonetheless offer inconsistent responses when probed with the same question, if separated by many other questions during which time they would presumably forget their earlier answers (for a review see Luce and Suppes 1965, §5).

More recently, Vul and Pashler (2008) performed a similar study, specifically on point estimation problems like the Africa/UN example above (§9), or the lifespace prediction problem (§3), in order to test whether averaging multiple guesses by the same subject would be more accurate than the average accuracy of their guesses, as has been demonstrated for groups of subjects (first, incidentally, by Galton 1889). They found this to be the case. Furthermore, averages were more accurate for subjects tested three weeks apart than twice on the same day, suggesting that samples become more independent as time passes. Thus, not only do subjects exhibit random variation, but responses appear to originate from some underlying distribution, which itself may encode more accurate knowledge of the world than any single sample drawn from it. In some sense this latter point is obvious. When making an estimate, say, about the percentage of African countries in the UN, we bring to bear all sorts of knowledge about Africa, the UN, and any other relevant topic, which cannot be captured by a single numerical estimate (for that matter, even if the point estimate were "optimal"). What is interesting about Vul and Pashler's results is that repeated elicitation of estimates gives evidence that subjects' intuitive theories of many of these domains are surprisingly accurate, and that these intuitive theories are organized in such a way as to produce samples from a sensible distribution, as the Sampling Hypothesis proposes.

11 Challenges

We have now reviewed several different sources of evidence for the Sampling Hypothesis: low-level perceptual and linguistic cognition and brain dynamics, characteristic sampling patterns of heuristics and biases in prediction, and random variation in individuals and populations suggestive of probabilistic sampling at the individual level. The number of papers exploring sample-based approximations to Bayesian models in particular has grown rapidly in the last several years, with work employing a variety of different algorithms, for a variety of psychological phenomena. We believe the summary given so far highlights the most important ideas and styles of application.

It should also be clear by now how the Sampling Hypothesis answers the three primary challenges to the idea that there could be probabilistic representation in the mind. On the issue of complexity, while it may be intractable to carry out arbitrary probabilistic calculations, drawing one or two (possibly biased) samples from a distribution only implicitly represented need not be very computationally intensive. It may be no more complex (possibly less) than simply imagining a few possible scenarios. On the issue of precision, recall the two considerations motivating the worry: that people do not "feel" as though they have precise probabilistic attitudes, and that it seems difficult, if not impossible, to measure mental states to such a level of precision. Since subjective probabilities are identified with certain physical propensities, the latter issue, while pressing, is no different from any other case in science of a "black-box" system that is difficult to measure directly. As for the former issue, we might expect a subject herself to be even less aware of her mind's own sampling propensities than a scientist trying to measure it. For instance, in the experiments described in Luce and Suppes (1965), and by Vul and Pashler, subjects were not even aware that

they were giving different answers to the same question. Finally, on the issue of coherence, as we have seen, the hypothesis is consistent with rather drastic deviations from consistency and coherence, given the inherent variability of sampling. Even if in any given case prediction is driven by a probabilistic mechanism, whose dynamics are supposed to represent subjective probabilities implicitly, this does not mean that subjects will always use these mechanisms in consistent ways, or in ways that produce behavior which we (from the outside) would deem probabilistically coherent.

Work on sampling has generated a lot of excitement in cognitive science, and deservedly so. But there are some difficult methodological and empirical issues that arise if we aim to gain a systematic understanding of when and how the mind makes use of sampling. We have already touched on some of them, but it is worth expanding on a few especially challenging themes. The central challenge is one of measurement: how do we go from a subject's (or a group of subjects') verbal and other behavior to a hypothesis about an internal sampleable generative model?

The standard way of proceeding in much of the psychological work discussed above, within a broadly Bayesian framework, is to hypothesize a sample space $\mathcal{V}$ and a "reasonable" distribution P, and compare subjects' responses with this distribution. P is typically conditioned on an event over $\mathcal{V}$ intended to correspond to a stimulus presented to subjects. As mentioned above, it is hypothesized that responses on the whole will more closely resemble the "optimal" response, given the model and the task, when stakes are high, or when the task is especially simple. When stakes are lower and the task is complex, subjects should tend to probability (or posterior) match (Vul et al., 2014). Given this, it ought even be possible to *predict* (the distribution of) individual responses, given a hypothesized model and an assumption about how subjects view the stakes.

This seems straightforward enough. In particular, the strategy seems no more difficult to carry out than any other type of Bayesian psychological modeling;[27] the only additional ingredient to be determined is the *number* of samples, which ought to be derived from some combination of stakes and complexity. Unfortunately the situation is not quite as simple as this description makes it appear. In addition to the usual challenges of determining the underlying representations, etc., we can identify at least three complicating factors particular to the Sampling Hypothesis.

The first is relatively minor, and has already been explored rather thoroughly in the literature. Drawing "perfect" samples from a distribution is in general as hard as computing the distribution explicitly. Consequently concrete algorithms for (cheaply) producing (biased) samples must be considered. Recall, for instance, that the Boltzmann Machine (Fig. 2) only produces a bona fide sample from the underlying distribution after it has run for a sufficient number of steps. Many concrete sampling algorithms have been explored in the statistics literature, and a fair number of these algorithms have been explored as potential psychological "process models." This plethora of possibilities can be seen as an advantage, where different psychological functions may call for different types of algorithms. But it also makes for a significantly greater degree of freedom in fitting models to data, since these algorithms would introduce bias into agents' predictive behavior in different ways.

A second, related, and perhaps more serious, conceptual concern is that some of these potential sampling algorithms may introduce bias in such a way that it no longer makes sense to construe the sampling propensities themselves as representing the agent's subjective uncertainty. To illustrate this, let us take an actual example of a recent proposal from Lieder et al. (2014), of so called *utility-weighted sampling*.[28] Without going into the technical details, the basic idea is easy to motivate. Suppose a person is trying to decide whether to approach a snake on a hiking trail. We might suppose that poisonous snakes are relatively rare in this location, so approaching it is likely to be safe. If the person were to use something like DECISION RULE B, we might expect her to sample only situations in which the snake is harmless. But clearly, for most of us the first thought we have in such a situation is that this snake might be dangerous. That is so even if on reflection we would assign low probability to the snake being dangerous. The suggestion in Lieder et al. (2014) is that our minds *overweight* extreme outcomes in the sampling process—sampling them at a rate much higher than their purported probability—in order to ensure that such extreme outcomes, which may be rare but also either disastrous or exceedingly positive, are taken into account in the decision making process.[29]

This work marks another interesting application of sampling to explain psychological biases (in this case, a version of the availability heuristic; see Lieder et al. 2014). But it raises yet further puzzles about how all of this work is

[27] As a number of authors have pointed out (e.g., Perfors 2012), demanding clarity about "where the priors come from"—what inductive biases people have—and "what the hypothesis space is"—what are the possible representations that could be used—ought to be seen as virtuous.

[28] Any concrete sampling algorithm would work as an illustration, but utility-weighted sampling illustrates the point especially vividly.

[29] To see what a difference this can make, Lieder et al. estimate that, in deciding whether to text and drive, if people were simply drawing samples in a way that mirrors the actual (objective) frequency, in order to have a 50% chance of conjuring to mind possible disaster scenarios, they would need to draw 700 million samples. By contrast, a single utility-weighted sample would have well over 99% chance of returning such a possibility.

supposed to fit together to form a coherent view of subjective probability. If subjective probabilities are identified with the sampling propensities in question, then what does it mean to say that they are *overweighted* in the sampling process? Overweighted with respect to what? Are the "true" subjective probabilities rather to be identified with some other mental construct whose role is always warped by utility assessments in the sampling process?

Perhaps a more reasonable interpretation of this work is that the "utility-weighted" sampling mechanism really does represent the subject's state of uncertainty as such. That is, what to us (and maybe to the subject herself, upon reflection) looks like a low probability event given the evidence is, for all unconscious intents and purposes, being treated by the mind as having high probability. Recalling part (C) of our characterization of subjective probability, whatever plays the role of probabilities should partly guide the agent's actions, in a way that makes sense given her probabilities and utilities. Evidently these "utility-weighted probabilities" do play that role: after all, for most people, seeing a snake, even from a distance, invokes a fright response that we may judge to be inappropriate upon reflection, given the statistics and the stakes of the situation.[30] Such a view coheres with the common observation that people overweight extreme hypotheses even in their explicit and conscious (but perhaps untutored) judgments (Lichtenstein et al., 1978).[31] So perhaps, counterintuitive though it may at first seem, such sampling propensities should in fact be seen as representing the agent's uncertainty.

Still, there are other puzzles in the vicinity. For instance, it seems that in some cases, what comes to mind neither looks like it should play a role in encoding the agent's uncertainty, nor does it apparently play any role in action. Even if by some quirk I keep imagining the snake all of a sudden reciting Shakespeare sonnets, this will presumably not guide my behavior in any straightforward way.[32] How can we distinguish the "imaginings" and other mental tokenings of events that do and do not play the role of a genuine sample from a sensical distribution?

Finally, a third challenge to the Sampling Hypothesis is the simple observation that very often our judgments would appear to be formed almost totally deterministically, even in the presence of evident uncertainty. For instance, suppose I present you with a six-sided dice, red painted on five sides, green on the sixth, and ask you to guess which color the next toss will turn up. Any minimally educated adult knows that there is a $1/6$ chance of green, but will undoubtedly guess red with probability 1, not merely $5/6$. Clearly people are not drawing a single sample from an internal model mirroring the dice propensities and then using DECISION RULE A. Even if they were drawing three samples, we would expect a green response once in about fifty trials ($1/54$), which still seems unlikely. It is too obvious in this case that red is the best guess.[33] This seems like another case where the mind overrides the results of any sampling mechanism, or perhaps does not invoke sampling at all.

As discussed earlier, it should be clear that evidence of sampling is not to be expected in every case of decision making. Yet, in order for the hypothesis to be more than haphazard observations about sundry phenomena, we would like some idea of when we should expect sampling to be operative in a given case. From the experiments reported in Luce and Suppes (1965) and Vul and Pashler (2008), we know that when people remember their earlier responses, their default sampling strategy is overridden (perhaps out of concern for consistency). For "obvious" problems like the dice example above, people also seem to invoke a more deterministic procedure. And, as mentioned in §6, my non-expert judgment that P $\neq$ NP is unlikely to result from any internal model I have about this domain, so *a fortiori* is unlikely to result from any sampling process over such a model. This gives at least three[34] examples of situations in which people do *not* seem to base their decisions and actions (at least purely) on the outcomes of an internal sampling

[30] On the view sketched here, such sampling propensities might be seen as a quantitative version of Gendler's *aliefs*, belief-like attitudes that elude rational control, but play an important role in quasi-automatic "reflex-like" behavior (Gendler, 2008). Thus, their role can be overridden by conscious control, as, e.g., when the snake is behind a pane of glass in a museum.

[31] Channeling the attitude expressed by Good (recall §3), we might construe the purpose of inductive *logic* precisely as overcoming such biases in how our naïve probability judgments (now construed as sampling propensities) are formed, so that our considered judgments about likelihood can be more purely based in evidence.

[32] However, intriguingly, it has recently been argued that the mere consideration of a proposition might amount to raising one's credence in it, if only very momentarily (Mandelbaum, 2014).

[33] This is assuming we have only a single trail. Empirically, when subjects make multiple sequential guesses, they sometimes do probability match, even when the "objective" probabilities are completely clear, as in this case (Vulcan, 2000).

One might observe something closer to sampling behavior in a slightly more complex scenario, where, say, the subject must make a choice between receiving a moderate-sized cash reward for sure, or a gamble that returns nothing on red and an enormous cash reward on green. In fact, these types of problems have been analyzed by Lieder et al. (2014) using utility-weighted sampling. See also Stewart et al. (2006) for a closely related decision-making paradigm based on sampling from memory traces.

[34] A fourth, quite curious, example comes from the observation that in many cases where adult subjects do probability match, young children seem to maximize robustly. See Yurovsky et al. (2013) for an overview and discussion.

process.

Judging from what we have seen in this paper, we might expect sampling to be in effect (very roughly) when the mind is faced with a *sufficiently difficult, but not excessively difficult* problem involving uncertainty, that can be "solved" unconsciously by invoking a *sufficiently rich* intuitive model of the domain. As it stands, this characterization is of course still too vague. Filling in the details is a significant challenge.

12 Further Comparison with Traditional Conceptions

In addition to presenting the Sampling Hypothesis as we see it, one of the aims of this paper is to understand how the resulting picture of subjective probability relates to traditional conceptions. As discussed earlier (§3), the view of subjective probability as sampling propensity is not intended to replace other notions of subjective probability. It is a hypothesis about one important way that our minds evidently come to *intuitive judgments* about situations under uncertainty. To be sure, these intuitive judgments can be overridden, and presumably should be in many cases, by conscious consideration and calculation. We can, if we like, follow Good in maintaining that, because our intuitive judgments often steer us wrong, we need more explicit probabilistic reasoning—including calculation with explicit numerical values—to correct for consequential biases and inaccuracies, which we fully expect on the sampling picture. This explicit reasoning might thus involve subjective probabilities in Good's sense (or one of the other familiar senses).

At this point, having seen what the Sampling Hypothesis amounts to, one might agree that it is an interesting psychological hypothesis worth taking seriously, and that it may be relevant to how we form more proper probability judgments, but nonetheless object to the claim that it offers a genuine notion of *subjective probability*. Perhaps subjective probabilities are supposed to play some additional theoretical role, on top of (A)-(C) offered in §2, and sampling propensities do not play this role. Here we consider two possible objections to this effect.

First, as already mentioned, the sampling propensities in question are quite removed from person-level phenomena, and in particular are inaccessible to introspective access or direct control. This applies most obviously to the low-level cognitive phenomena we discussed, such as parsing and vision, but it seems to apply just as well to the high-level phenomena. For instance, in estimating the number of African countries in the UN, if what people are doing is running a Markov chain from some initial salient value, they certainly are not aware of this. Does that disqualify the underlying probabilities as being representations of *the person's own* subjective probabilities?

If it did, it would presumably disqualify several other familiar accounts as well. When we identify a person's subjective probabilities with values inferred from an expected-utility representation of their qualitative preferences, assuming those preference obey certain axioms (e.g., as in Savage 1954), it is no more apparent that the resulting probabilities are *the person's own* in this robust sense. The same goes for the interpretive account, according to which these probabilities are similarly assigned "from outside" (Lewis, 1974). In all of these cases, when asked directly how likely she takes some event to be, a subject may be unable to come up with a numerical estimate; or else her own estimate may be quite different from what is ascribed to her by a third party on the basis of her choices or actions. Recall that Ramsey was famously adamant that a theory of subjective probability could not be based on "introspectible feelings" (Ramsey, 1931). Instead, the study should be tied to some "causal properties" of the subjective state. For Ramsey and many following, the link is directly and inextricably to action. For us, the "causal property" is a hypothesized (and in principle observable) internal state of the agent's brain. Either way, these states are not the sort of thing that an agent must be able to introspect. Moreover, as discussed in §2, we want a notion that can be applied all the way down to low-level perception, parsing, and so on.

However, there is a second, related concern, which bears on the potential philosophical interest of subjective probability. Namely, does this species of probability play the same *normative* role that subjective probability is supposed to play? Assume for the moment that we do not want to analyze subjective probabilities in terms of representation theorems or external rationalizations, but instead take the notion to be an unanalyzed conceptual primitive (see, e.g., Eriksson and Hájek 2007 for this line). A common view is that human agents have graded attitudes, called *degrees of belief* or *credences*, and that, for one reason or another, these graded attitudes *ought* to conform to the axioms of probability. For instance, Dutch book theorems (de Finetti, 1974), accuracy-dominance theorems (Joyce, 1998), and axiomatic derivations (Jaynes, 2003) are some of the tools used to support common arguments for this normative conclusion. The philosophy of subjective probability, on this view, is about norms for degrees of belief. By contrast, our subjective probabilities (as explained in §7) satisfy the probability axioms just by virtue of being propensities, defined

over a space of return values. There is no (normative) question about whether these probabilities ought to satisfy the axioms; they just do. Does this show there are really two different subject matters here?

We believe not. Consider how philosophers often introduce the subject matter of degrees of belief. Simplifying somewhat, a typical paper will begin with some platitudes about believing some things more than others ("to a greater degree") and argue that we need a theory about which such "graded beliefs" could be rational (rationally held together, etc.). The following introductory passage from Christensen (1996) is representative:

> A little reflection on my beliefs is enough to convince me that they come in degrees. I believe that I shall eat Italian sausages tonight, and I believe that the sun will set tonight, but I am much less confident about the sausages than about the sunset. (450)

Note that these are quite ordinary beliefs, not unlike the belief that it will rain tomorrow, or that there are 50 African countries in the UN, or that an arbitrarily chosen room with a lamp will also contain a chair. The mere fact that some of these beliefs more vividly compel us, or that we can judge some to hold more strongly than others, does not by itself show that the states in question can be meaningfully assigned numerical quantities (or even sets of numerical quantities).[35] On the traditional picture—where the subjective probabilities are constructed, as for Ramsey, de Finetti, Savage, etc.—numerical talk is legitimized by linking the states to something independent, such as betting behavior or preferences. Such numerical talk may also not need legitimization when probabilities are expressed explicitly by the subject. But if we maintain that tacit subjective probabilities are a central part of a person's psychology, operative in stock examples, and we do not want to *reduce* them to betting or choice behavior—and for instance, want to take them as primitive—we are left with the question of what these numbers mean.

The Sampling Hypothesis offers a compelling, empirically grounded, neurally plausible story about what these numbers could mean. As we have been attempting to show, sampling propensities *are* probabilities, hence are already numerical, they can meaningfully be said to represent the agent's own uncertainty, they are typically responsive to some sources of data, and they play an appropriate role in guiding (at least certain aspects of) the agent's behavior. To the extent that sampling will turn out to be central in explanations of much of our prediction and decision-making behavior, one might argue that what we really want to understand are norms appropriate for agents that look like this.

What would such norms look like? As already discussed, questions about consistency and coherence cannot be questions about a person's subjective probabilities, if those amount to sampling propensities. Rather, such issues may be folded into more general questions, for instance, of when a subject ought to use which internal generative model for sampling (or perhaps to use some other non-sampling-based mechanism). If we want to say that a *mistake* is being made when a person declares that a word is more likely to end in 'ing' than in '-n-', the natural way to say this is not that her degrees of belief conflict with the probability axioms, but that she is using her mental toolbox to answer the question in the wrong way. She might be using a procedure for answering the question, which may be prudent in general—specifically, we suggested it might be useful when the alternative possibilities are disparate—but produces logically contradictory results in this case. Indeed, we should expect this kind of *fragmentation* to be the rule rather than the exception: in addition to expressions of incoherence or inconsistency, the processes in question will also be amenable to context and framing effects, making these subjective probabilities variable over time and contexts, rather than being a permanent, stable state of a person.[36]

The view of subjective probability as sampling propensity generally opens up natural questions pertaining to *bounded* or *procedural rationality* (Simon, 1976) in the context of reasoning under uncertainty. For instance, on the assumption that drawing a sample takes time and energy, and thus comes with a cost, one can ask the question: given a decision problem, a sampling mechanism, and a cost for drawing each additional sample, what is the *best* way of using that mechanism to solve the decision problem under resource constraints? This type of question has now been investigated under numerous guises, including issues of how many samples to draw (Vul et al., 2014), how much bias to tolerate in drawing samples (Lieder et al., 2012), and which nodes in a graphical model to "ignore" to facilitate faster sampling (Icard and Goodman, 2015). This style of work allows asking sharper questions about rationality,

[35] Indeed, a number of theorists have claimed that numbers cannot be meaningfully assigned to such states. We already mentioned Harman. Keynes (1921) is another famous example, and of course there are many others, in philosophy, psychology, and the social sciences.

[36] In this sense, the Sampling Hypothesis seems to be complementary to a view expressed recently by Norby (2015), who claims that degrees of belief as traditionally conceived are inappropriate as descriptions of ordinary psychological states that feature in decision making. Interestingly, Norby comes close to expressing something in the neighborhood of the Sampling Hypothesis (see his §3 on "proto-credences"), based on some of the same empirical work (in particular Stewart et al. 2006). However, he does not takes these states to be genuine subject probabilities (87).

taking resource limitations into account, which also has the potential virtue of coming closer to concrete questions about how beings like us, who certainly share such limitations, ought to think and behave.[37] In the other direction, it has been suggested that such (bounded) rationality considerations might play an important methodological role in hypothesis formation about how various cognitive mechanisms work (Griffiths et al., 2015; Icard, 2014), echoing earlier suggestions by both philosophers and psychologists (Dennett, 1981; Anderson, 1990).

To repeat once again, this is not to deny that there are other kinds of normative questions one can ask, e.g., that might apply to betting odds or other species of subjective probability. However, rather than casting doubt on whether sampling propensities ought to be considered *bona fide* subjective probabilities, we ought to see these new issues as important reflections of the fact that, however probabilities are represented in the mind (to the extent that they are), they must be represented in some format, and thus will inevitably require computational resources to use effectively. The Sampling Hypothesis provides a concrete proposal concerning that format, and thus presents a chance for asking novel and precise questions about rationality in the context of probabilistic inference and reasoning.

13 Conclusion

According to the Sampling Hypothesis, much of what we pre-theoretically think of as *graded belief*, and much of what goes under the heading of *subjective probability* in theorizing about human agency and intelligence, is represented in the brain as a kind of *sampling propensity* of an internal generative model. In this way an important variety of subjective probability is reduced to a kind of objective probability, assumed to be operative in our thought and decision making via a choice mechanism like DECISION RULE A or B. We have tried to present the view as clearly as possible, both at a conceptual level and at an empirical level, offering the most compelling evidence in its favor, and to show that the view overcomes many of the common challenges to the idea that the mind does in fact use probabilities.

The Sampling Hypothesis ties together a number of otherwise disparate ideas and traditions within cognitive science into a single view: probabilistic models of learning and reasoning, connectionist models, the stochastic nature of choice and behavior, heuristics and biases, multistability in perception, probability matching, and bounded rationality. We also identified a number of conceptual and empirical challenges for the hypothesis. Given the nascent but promising empirical support for the hypothesis, and its potential for clarifying the nature of philosophically important mental states and processes, raising new normative questions in addition, we believe the Sampling Hypothesis merits further attention, experimental investigation, and systematization.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Lawrence Earlbaum Associates, Inc.

Anscombe, F. J. and Aumann, R. J. (1963). A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205.

Arora, S. and Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–609.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331:83–87.

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7(11).

Carnap, R. (1947). On the application of inductive logic. *Philosophy and Phenomenological Research*, 8:133–148.

Chater, N. and Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Science*, 10(7):335–344.

[37] As mentioned above (§9), it has been common in this literature to rationalize behavior at odds with logic or probability by invoking notions of boundedness or resource constraints.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Science*, 10(7):287–291.

Christensen, D. (1996). Dutch-book arguments depragmatized: Epistemic consistency for partial believers. *Journal of Philosophy*, 93:450–479.

Churchland, P. S. and Sejnowski, T. J. (1994). *The Computational Brain*. MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36:181–253.

Craik, K. (1943). *The Nature of Explanation*. Cambridge University Press.

Davidson, D. (1975). Hempel on explaining action. *Erkenntnis*, 10(3):239–253.

Denison, S., Bonawitz, E., Gopnik, A., and Griffiths, T. L. (2013). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition*, 126:285–300.

Dennett, D. C. (1981). Three kinds of intentional psychology. In Healey, R., editor, *Reduction, Time, and Reality*, pages 37–61. Cambridge University Press.

Dreyfus, H. L. and Dreyfus, S. E. (1986). *Mind over Machine*. Free Press.

Eriksson, L. and Hájek, A. (2007). What are degrees of belief? *Studia Logica*, 86(2):183–213.

de Finetti, B. (1974). *Theory of Probability*, volume 1. Wiley, New York.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Science*, 14(3):119–130.

Freer, C., Roy, D., and Tenenbaum, J. (2012). Towards common-sense reasoning via conditional simulation: Legacies of Turing in artificial intelligence. In Downey, R., editor, *Turing's Legacy*. ASL Lecture Notes in Logic.

Gaifman, H. (2004). Reasoning with limited resources and assigning probabilities to arithmetical statements. *Synthese*, 140:97–119.

Galton, F. (1889). *Natural Inheritance*. MacMillan.

Gendler, T. (2008). Alief and belief. *Journal of Philosophy*, 105(10):634–663.

Gershman, S. J. and Daw, N. D. (2012). Perception, action, and utility: the tangled skein. In Rabinovich, M., Friston, K., and Varona, P., editors, *Principles of Brain Dynamics: Global State Interactions*, pages 293–312. MIT Press.

Gershman, S. J., Vul, E., and Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, 24:1–24.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–699.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press.

Goodman, N. D., Tenenbaum, J. B., and Gerstenberg, T. (2014). Concepts in a probabilistic language of thought. In Margolis, E. and Laurence, S., editors, *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1):3–32.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Science*, 14(8):357–364.

Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2):217–229.

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

Harman, G. (1986). *Change in View*. MIT Press.

Icard, T. (2013). *The Algorithmic Mind: A Study of Inference in Action*. PhD thesis, Stanford University.

Icard, T. F. (2014). Toward boundedly rational analysis. In Bello, P., Guarini, M., McShane, M., and Scassellati, B., editors, *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 637–642.

Icard, T. F. and Goodman, N. D. (2015). A resource-rational approach to the causal frame problem. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., and Maglio, P. P., editors, *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

James, W. (1890). *The Principles of Psychology*. Henry Holt & Co.

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

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65:575–603.

Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Kruschke, J. K. (2006). Locally Bayesian learning with application to retrospective revaluation and highlighting. *Psychological Review*, 113(4):677–699.

Kwisthout, J., Wareham, T., and van Rooij, I. (2008). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, 35:779–784.

Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America, A*, 20(7):1434–1448.

Levy, R., Reali, F., and Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in Neural Information Processing Systems*, 21:937–944.

Lewis, D. K. (1974). Radical interpretation. *Synthese*, 23:331–344.

Lichtenstein, S., Slovic, P., Fischoff, B., Layman, M., and Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6).

Lieder, F., Griffiths, T. L., and Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems*, 25:2699–2707.

Lieder, F., Hsu, M., and Griffiths, T. L. (2014). The high availability of extreme events serves resource-rational decision-making. In Bello, P., Guarini, M., McShane, M., and Scassellati, B., editors, *Proceedings of the 36th Annual Meeting in Cognitive Science*.

Lochmann, T. and Deneve, S. (2011). Neural processing as causal inference. *Current Opinion in Neurobiology*, 21:774–781.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons.

Luce, R. D. and Suppes, P. (1965). Preference, utility, and subjective probability. In Luce, R. D., Bush, R. R., and Galanter, E. H., editors, *Handbook of Mathematical Psychology*, volume 3, pages 249–410. Wiley.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mandelbaum, E. (2014). Thinking is believing. *Inquiry: An Interdisciplinary Journal of Philosophy*, 57(1):55–96.

Marr, D. (1982). *Vision*. W.H. Freeman and Company.

McFadden, D. L. (1973). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*. Academic Press.

Millgram, E. (1991). Harman's hardness arguments. *Pacific Philosophical Quarterly*, 72(3):181–202.

Moreno-Bote, R., Knill, D. C., and Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30):12491–12496.

Mozer, M. C., Pashler, H., and Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, 32:1133–1147.

Norby, A. (2015). Uncertainty without all the doubt. *Mind and Language*, 30(1):70–94.

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

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

Raiffa, H. (1968). *Decision Analysis*. Addison-Wesley.

Ramsey, F. P. (1931). Truth and probability. In Braithwaite, R. B., editor, *Foundations of Mathematics and Other Logical Essays*. Martino Fine.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Rumelhart, D. E., McClelland, J. L., and The PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons.

Schacter, D. L., Addis, D. R., and Buckner, R. L. (2008). Episodic simulation of future events: Concepts, data, and applications. *Annals of the New York Academy of Sciences*, 1124:39–60.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., and Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, 61(2):195–202.

Seth, A. K. (1999). Evolving behavioural choice: An exploration of Hernnstein's Matching Law. In Floreano, D., Nicoud, J.-D., and Mondada, F., editors, *Proceedings of the Fifth European Conference on Artificial Life*, pages 225–236. Springer.

Shi, L., Griffiths, T. L., Feldman, N. H., and Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, 17(4):443–464.

Simon, H. A. (1976). From substantive to procedural rationality. In Kastelein, T. J., Kuipers, S. K., Nijenhuis, W. A., and Wagenaar, G. R., editors, *25 Years of Economic Theory*, pages 65–86. Springer.

Solway, A. and Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference. *Psychological Review*, 119(1):120–154.

Stewart, N., Chater, N., and Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, 53:1–26.

Sundareswara, R. and Schrater, P. (2007). Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision*, 8(5):1–19.

Suppes, P. (1974). The measurement of belief. *The Journal of the Royal Statistical Society, Series B*, 36(2):160–191.

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

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315.

Vilares, I. and Kording, K. P. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, 1224:22–39.

Vul, E. (2010). *Sampling in Human Cognition*. PhD thesis, MIT.

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

Vul, E. and Pashler, H. (2008). Measuring the crowd within. *Psychological Science*, 19(7):645–647.

Vulcan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, 13(1):101–118.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall.

Williamson, T. (2015). Acting on knowledge. In Carter, J. A., Gordon, E., and Jarvis, B., editors, *Knowledge-First*. Oxford University Press.

Yang, T. and Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, 447:1075–1082.

Yurovsky, D., Boyer, T. W., Smith, L. B., and Yu, C. (2013). Probabilistic cue combination: Less is more. *Developmental Science*, 16(2):149–158.

Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science*, 67(1):45–69.

A Boltzmann Machine as Gibbs Sampler

In this Appendix we explain the sense in which the activation rule for the Boltzmann Machine carries out Gibbs sampling on an underlying Markov random field. This fact is folklore in cognitive science.

Gibbs sampling is an instance of the Metropolis-Hastings algorithm, which in turn is a typle of Markov chain Monte Carlo inference (MacKay, 2003). Suppose we have a multivariate distribution $P(X_1,\ldots,X_n)$ and we want to draw samples from it. The Gibbs sampling algorithm is as follows:

1. Specify some initial values for all the random variables $\mathbf{y}^{(0)} = (y_1^{(0)},\ldots,y_n^{(0)})$.

2. Given $\mathbf{y}^{(r)}$, randomly choose a number $i \in \{0,\ldots,n\}$, and let $\mathbf{y}^{(r+1)}$ be exactly like $\mathbf{y}^{(r)}$, except that $y_i^{(r+1)}$ is redrawn from the conditional distribution $P(X_i \mid y_{-i})$.

3. At some stage R, return some subset of $\{\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(R)}\}$ as samples.

Note that the sequence of value vectors $\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(r)}, \ldots$ forms a Markov chain because the next sample $\mathbf{y}^{(r+1)}$ only depends on the previous sample $\mathbf{y}^{(r)}$. Let $q(\mathbf{y} \to \mathbf{y}')$ be the probability of moving from $\mathbf{y}$ at a given stage r to $\mathbf{y}'$ at stage $r+1$ (which is the same for all r, and equal to 0 when $\mathbf{y}$ and $\mathbf{y}'$ differ by more than one coordinate.). And let $\pi^r(\mathbf{y})$ be the probability of being in state $\mathbf{y}$ at stage r. Then we clearly have:

$$\pi^{r+1}(\mathbf{y}') = \sum_{\mathbf{y}} \pi^r(\mathbf{y}) q(\mathbf{y} \to \mathbf{y}') \; .$$

By general facts about Markov chains, it can be shown that this process reaches a unique stationary distribution, i.e., $\pi^r = \pi^{r+1}$ for all $r > s$ for some s. This is the unique distribution π^* for which the following holds:

$$\pi^*(\mathbf{y}') = \sum_{\mathbf{y}} \pi^*(\mathbf{y}) q(\mathbf{y} \to \mathbf{y}') \; .$$

Since this equation also holds for P, that shows the Markov chain converges to $P = \pi^*$.

Recall the Boltzmann Machine is defined by a set of nodes and weights between those nodes.[38] If we think of the nodes as binary random variables $(X_1, \ldots, X_n)$, taking on values $\{0,1\}$, then the weight matrix $\mathbf{W}$ gives us a natural distribution $P(X_1, \ldots, X_n)$ on state space $\{0,1\}^n$ as follows. First, define an *energy function* E on the state space:

$$E(\mathbf{y}) = -\frac{1}{2} \sum_{i,j} W_{i,j} y_i y_j \; .$$

Then the resulting *Boltzmann distribution* is given by:

$$P(\mathbf{y}) = \frac{e^{-E(\mathbf{y})}}{\sum_{\mathbf{y}'} e^{-E(\mathbf{y}')}} \; .$$

Note that the denominator becomes a very large sum very quickly. Recall from our earlier discussion that the size of the state space is 32 with 5 nodes, but over a trillion with 40 nodes. Suppose we apply the Gibbs sampling algorithm to this distribution. Then at a given stage, we randomly choose a node to update, and the new value is determined by the conditional probability as above. This step is equivalent to applying the activation function for the Boltzmann Machine (where in the following, $\mathbf{y}'$ is just like $\mathbf{y}$, except that $y'_i = 1$ and $y_i = 0$):

$$\begin{aligned}
P(X_i = 1 \mid \{y_j\}_{j \neq i}) &= P(X_i = 1 \mid \mathbf{y} \text{ or } \mathbf{y}') \\
&= \frac{P(\mathbf{y}')}{P(\mathbf{y} \text{ or } \mathbf{y}')} \\
&= \frac{e^{-E(\mathbf{y}')}}{e^{-E(\mathbf{y})} + e^{-E(\mathbf{y}')}} \\
&= \frac{1}{1 + e^{E(\mathbf{y}') - E(\mathbf{y})}} \\
&= \frac{1}{1 + e^{-\sum_j W_{i,j} y_j}} \\
&= \frac{1}{1 + e^{-net_i}} \; .
\end{aligned}$$

Thus, the above argument shows that the Boltzmann Machine (eventually) samples from the associated Boltzmann distribution. Incidentally, we can also see now why net_i is equivalent to the log odds ratio under the Boltzmann

[38] We ignore bias terms here to simplify the presentation. See Rumelhart et al. (1986) for the general formulation.

distribution (recall the discussion in §8 of the Yang and Shadlen 2007 experiment):

$$\begin{aligned}\frac{P(\mathbf{y}')}{P(\mathbf{y})} &= \frac{e^{-E(\mathbf{y}')}}{e^{-E(\mathbf{y})}} \\ &= e^{E(\mathbf{y})-E(\mathbf{y}')} \\ &= e^{\sum_j W_{i,j} y_j} \\ &= e^{net_i} .\end{aligned}$$

And thus,

$$\begin{aligned}\log \frac{P(\mathbf{y}')}{P(\mathbf{y})} &= \log\left(e^{net_i}\right) \\ &= net_i .\end{aligned}$$

B Softmax versus Sampling Rule

In this Appendix, we offer some observations about the relation between the softmax (or generalized Luce-Shepard) choice rule and the sampling-based decision rules discussed in this chapter. Suppose we associate with a subject a probability function $P(\cdot)$ on sample space $\mathcal{H} = \{H_1, \ldots, H_n\}$ and a utility u over actions $\mathcal{A} = \{A_1, \ldots, A_m\}$.

The softmax rule says that the subject will give response A with probability

$$\frac{e^{v(A)/\beta}}{\sum_{A' \in \mathcal{A}} e^{v(A')/\beta}},$$

where v is some value function, which in this case we will assume is the log expected utility:

$$v(H) = \log \sum_{H \in \mathcal{H}} P(H) u(A, H).$$

As a representative example of a sampling based rule, recall DECISION RULE B:

> DECISION RULE B: Suppose we are given a generative model $\mathcal{M}$ with $\mathcal{V}$ as possible return values. To select an action, take R samples, $H^{(1)}, \ldots, H^{(R)}$, using $\mathcal{M}$, and let BEST be the set of actions that receive the largest summed utilities, i.e.,
>
> $$\text{BEST} = \{A_j : \sum_{i=1}^{R} u(A_j, H^{(i)}) \text{ is maximal}\}.$$
>
> Take action $A_j \in$ BEST with probability $\frac{1}{|\text{BEST}|}$.

In the specific case of a certain kind of estimation problem—where $\mathcal{H} = \mathcal{A}$ and the utility of an estimate is 1 if correct, 0 otherwise; thus expected utility and probability coincide—it is easy to see that the softmax rule with $\beta = 1$ and DECISION RULE B with $R = 1$ (or $R = 2$) are equivalent. The probability of returning hypothesis H is just $P(H)$, i.e., we have probability matching.

Unfortunately, even restricting to these simple estimation problems, the relation between the two rules as functions of β and R is intractable and varies with the probability distribution $P(\cdot)$, as Vul (2010) points out. Thus, beyond $\beta = R = 1$ it is hard to study their relationship. As mentioned in the text, both can be used to fit much of the psychological data, though one might suspect the sampling rule is more reasonable on the basis of computational considerations.

Interestingly, for more general classes of decision problems, these two classes of rules can be qualitatively distinguished. The Luce Choice Axiom, from which the Luce choice rule was originally derived (Luce, 1959), gives a hint of how we might do this. Where $P_S(T)$ is the probability of choosing an action from $T \subseteq \mathcal{A}$ from among options in $S \subseteq \mathcal{A}$, the choice axiom states that for all R such that $T \subseteq R \subseteq S$:

$$P_S(T) \;=\; P_S(R)\, P_R(T) .$$

It is easy to see the softmax rule satisfies the choice axiom for all values of β:

$$\begin{aligned} P_S^{\text{softmax}}(T) &= \frac{\sum_{A\in T} e^{v(A)/\beta}}{\sum_{A\in S} e^{v(A)/\beta}} \\ &= \frac{\sum_{A\in R} e^{v(A)/\beta}}{\sum_{A\in S} e^{v(A)/\beta}} \cdot \frac{\sum_{A\in T} e^{v(A)/\beta}}{\sum_{A\in R} e^{v(A)/\beta}} \\ &= P_S^{\text{softmax}}(R)\, P_R^{\text{softmax}}(T)\,. \end{aligned}$$

For estimation problems, DECISION RULE B with $R = 1$ also satisfies this axiom. However, in more general contexts, even for the case of $R = 1$, it does not. Perhaps the simplest illustration of this is the decision problem in Table 1, where $\mathcal{H} = \{H_1, H_2\}$ and $\mathcal{A} = \{A_1, A_2, A_3\}$, and $\varepsilon > 0$.

	H_1	H_2
A_1	$1-\varepsilon$	$1-\varepsilon$
A_2	2	0
A_3	0	2

Table 1: Distinguishing softmax and sample-based decision rules

We clearly have $P^{\text{RULE B}}_{\{A_1,A_2,A_3\}}(\{A_1\}) = 0$. Yet, as long as $P(H_1), P(H_2) > 0$ and $\varepsilon < 1$, we have

$$P^{\text{RULE B}}_{\{A_1,A_2,A_3\}}(\{A_1,A_2\})\, P^{\text{RULE B}}_{\{A_1,A_2\}}(\{A_1\}) > 0\,,$$

showing the violation of the choice axiom. As the softmax rule satisfies the axiom, this gives us an instance where we would expect different behavior, depending on which rule (if either) a subject is using. When presented with such a problem, a softmax-based agent would sometimes choose action A_1. In fact, the probability of choosing A_1 can be nearly as high as $1/3$, for example, when $\beta = 1$, ε is very small, and H_1 and H_2 are equiprobable. However, for any set of samples of any length R, one of A_2 or A_3 would always look preferable for a sampling agent. Thus, such an agent would never choose A_1. It would be interesting to test this difference experimentally. Doing so could offer important evidence about the tenability of the Sampling Hypothesis.