Cognitive Psychology 96 (2017) 1–25

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Where do hypotheses come from?

Ishita Dasgupta [a,*], Eric Schulz [b], Samuel J. Gershman [c]

[a] *Department of Physics and Center for Brain Science, Harvard University, United States*
[b] *Department of Experimental Psychology, University College London, United Kingdom*
[c] *Department of Psychology and Center for Brain Science, Harvard University, United States*

ARTICLE INFO

Article history:
Accepted 16 May 2017
Available online 3 June 2017

Keywords:
Hypothesis generation
Bayesian inference
Monte Carlo methods

ABSTRACT

Why are human inferences sometimes remarkably close to the Bayesian ideal and other times systematically biased? In particular, why do humans make near-rational inferences in some natural domains where the candidate hypotheses are explicitly available, whereas tasks in similar domains requiring the self-generation of hypotheses produce systematic deviations from rational inference. We propose that these deviations arise from algorithmic processes approximating Bayes' rule. Specifically in our account, hypotheses are generated stochastically from a sampling process, such that the sampled hypotheses form a Monte Carlo approximation of the posterior. While this approximation will converge to the true posterior in the limit of infinite samples, we take a small number of samples as we expect that the number of samples humans take is limited. We show that this model recreates several well-documented experimental findings such as anchoring and adjustment, subadditivity, superadditivity, the crowd within as well as the self-generation effect, the weak evidence, and the dud alternative effects. We confirm the model's prediction that superadditivity and subadditivity can be induced within the same paradigm by manipulating the unpacking and typicality of hypotheses. We also partially confirm our model's prediction about the effect of time pressure and cognitive load on these effects.

1. Introduction

In his preface to *Astronomia Nova* (1609), Johannes Kepler described how he struggled to find an accurate mathematical description of planetary motion. Like most of his contemporaries, he started with the hypothesis that planets move in perfect circles. This necessitated extraordinary labor to reconcile the equations of motion with his other assumptions, "because I had bound them to millstones (as it were) of circularity, under the spell of common opinion." It was not the case that Kepler simply favored circles over ellipses (which he ultimately accepted), since he considered several other alternatives prior to ellipses. Kepler's problem was that he failed to generate the right hypothesis.[1]

Kepler is not alone: the history of science is replete with examples of "unconceived alternatives" (Stanford, 2010), and many psychological biases can be traced to failures of hypothesis generation, as we discuss below. In this paper, we focus

* Corresponding author at: Department of Physics and Center for Brain Science, Harvard University, 52 Oxford Street, Room 295.08, Cambridge, MA 02138, United States.

E-mail address: idasgupta@physics.harvard.edu (I. Dasgupta).

[1] In fact, Kepler had tried fitting an oval to his observations only to reject it, and then labored for another seven years before finally trying an ellipse and realizing that it was mathematically equivalent to an oval. As he recounted, "The truth of nature, which I had rejected and chased away, returned by stealth through the back door, disguising itself to be accepted... Ah, what a foolish bird I have been!".

http://dx.doi.org/10.1016/j.cogpsych.2017.05.001

on hypothesis generation in the extensively studied domain of probabilistic inference. The generated hypothesis are a subset of a tremendously large space of possibilities. Our goal is to understand how humans generate that subset.

In general, probabilistic inference is comprised of two steps: hypothesis generation and hypothesis evaluation, with feedback between these two processes. Given a complete set of hypotheses $\mathcal{H}$ and observed data d, optimal evaluation is prescribed by Bayes' rule, which assigns a posterior probability $P(h|d)$ to each hypothesis $h \in \mathcal{H}$ proportional to its prior probability $P(h)$ and the likelihood of the observed data under $h, P(d|h)$:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}. \tag{1}$$

Many studies have found that when $\mathcal{H}$ is supplied explicitly, humans can come close to the Bayesian ideal (e.g. Frank & Goodman, 2012; Griffiths & Tenenbaum, 2006, 2011; Oaksford & Chater, 2007; Petzschner, Glasauer, & Stephan, 2015).[2] However, when humans must generate the set of hypotheses themselves, they cannot generate them all and instead generate only a subset, leading to judgment biases (Carroll & Kemp, 2015; Dougherty & Hunter, 2003; Gettys & Fisher, 1979; Koriat, Lichtenstein, & Fischhoff, 1980; Thomas, Dougherty, Sprenger, & Harbison, 2008; Weber, Böckenholt, Hilton, & Wallace, 1993). Some prominent biases of this kind are listed in Table 1.

Most previously proposed models of hypothesis generation rely on cued recall from memory based on similarity to previously observed scenarios (c.f. Gennaioli & Shleifer, 2010; Thomas et al., 2008). The probability of a generated hypothesis depends on the strength of its memory, and the number of such hypotheses generated is constrained by the available working memory resources. However, in most naturally encountered combinatorial hypothesis spaces, the number of possible hypotheses is vast and only ever sparsely observed. Goodman, Tenenbaum, Feldman, and Griffiths (2008) showed that, when inferring Boolean concepts, people can generate previously unseen hypotheses by using compositional rules, instead of likening the situation to previously observed situations. So it seems that humans do not generate hypotheses only from the manageably small subset of previously observed hypotheses in memory and instead are able to generate hypotheses from the formidably large combinatorial space of all the conceivable possibilities. Given how large this space is, resource constraints at the time of inference suggest that only a subset are actually generated.

In this paper, we develop a normative theoretical framework for hypothesis generation in the domain of probabilistic inference, given fixed data, arguing that the brain copes with the intractability of inference by stochastically sampling hypotheses from the combinatorial space of possibilities (see also Sanborn & Chater, 2016). Although this sampling process is asymptotically exact, time pressure and cognitive resource constraints limit the number of samples that can be generated, giving rise to systematic biases. Such biases are "computationally rational" in the sense that they result from a trade-off between the costs and benefits of computation—i.e., they are an emergent property of the expected utility calculus when costs of computation are taken into account (Gershman, Horvitz, & Tenenbaum, 2015; Lieder, Griffiths, Huys, & Goodman, 2017a; Vul, Goodman, Griffiths, & Tenenbaum, 2014). We propose that the framing of a query leads to sampling specific hypotheses first, which biases the rest of the hypothesis generation process through correlations in the sampling process. We discuss the properties of various sampler designs to explore the space of possible algorithms, and choose a specific design that can reproduce all the phenomena listed in Table 1. We then test our theory's novel predictions in four experiments.

2. A rational process model of hypothesis generation

Much of the recent work on probabilistic inference in human cognition has been deliberately agnostic about its underlying mechanisms, in order to make claims specifically about the subjective probability models people use in different domains (Chater et al., 2006). Because the posterior distribution $P(h|d)$ is completely determined by the joint distribution $P(h,d) = P(d|h)P(h)$, an idealized reasoner's inferences can be perfectly predicted given this joint distribution. By comparing different assumptions about the joint distribution (e.g., the choice of prior or likelihood) under these idealized conditions, researchers have attempted to adjudicate between different models. Importantly, any algorithm that computes the exact posterior will yield identical predictions, which is what licenses agnosticism about mechanism. This method of abstraction is the essence of the "computational level of analysis" (Marr & Poggio, 1976), and is closely related to the competence/performance distinction in linguistics and "as-if" explanations of choice behavior in economics.

The phenomena listed in Table 1 do not yield easily to a purely computational-level analysis, since different choices for the probabilistic model do not account for the systematic errors in approximating them. For this reason, we turn to "rational process" models (see Griffiths, Vul, & Sanborn, 2012, for a review), which make explicit claims about the mechanistic implementation of inference. Rational process models are designed to be approximations of the idealized reasoner, but make distinctive predictions under resource constraints. In particular, we explore how sample-based approximations lead to particular cognitive biases in a large space of hypotheses, when the number of samples is limited. With an infinite number

[2] This correspondence between human and Bayesian inference requires that the inference task must be one that is likely to have been optimized by evolution (e.g., predicting the duration of everyday events, categorizing and locating objects in images, making causal inferences), typically in domains where people have strong intuitive knowledge about the relative probabilities of hypotheses; asking humans to reason consciously about unnatural problems like randomness or rare events (see Chater, Tenenbaum, & Yuille, 2006, for discussion), or carry out explicit updating calculations (Peterson & Beach, 1967), tends to produce deviations from the Bayesian ideal.