Cognition 128 (2013) 417–423

Contents lists available at SciVerse ScienceDirect

Cognition

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

Brief article

Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013)

Michael C. Frank *

Department of Psychology, Stanford University, United States

ARTICLE INFO

Article history:
Received 16 January 2013
Revised 29 April 2013
Accepted 30 April 2013
Available online 15 June 2013

Keywords:
Rule learning
Bayesian models
Cognitive modeling

ABSTRACT

A recent probabilistic model unified findings on sequential generalization ("rule learning") via independently-motivated principles of generalization (Frank and Tenenbaum, 2011). Endress critiques this work, arguing that learners do not prefer more specific hypotheses (a central assumption of the model), that "common-sense psychology" provides an adequate explanation of rule learning, and that Bayesian models imply incorrect optimality claims but can be fit to any pattern of data. Endress's response raises useful points about the importance of mechanistic explanation, but the specific critiques of our work are not supported. More broadly, I argue that Endress undervalues the importance of formal models. Although probabilistic models must meet a high standard to be used as evidence for optimality claims, they nevertheless provide a powerful framework for describing cognition.

1. Introduction

How do you reverse engineer an alien computer? Figuring out how it works requires moving back and forth between what you can learn about its individual parts and broader hypotheses about its function and governing principles. The general theory of computation leads to questions about the artifact's inputs, outputs, and methods for storing information (Hopcroft et al., 1979). But since computational systems can store their state in processes as diverse as symbols on a tape or weights between neurons (McCulloch and Pitts, 1943), a high-level understanding of the device provides only general constraints on lower-level hypotheses. In Marr's (1982) terms, a *computational* level understanding of the system needs to be integrated with both a model of the system's sub-components (the *algorithmic* level) and, critically, an understanding of the individual units of the system (the *implementational* level). Each of these levels of representation contributes to the ability to repair, duplicate, and extract general insights from the artifact.

Reverse engineering the human mind requires the same attention to multiple levels of abstraction. A wide range of theorists have recognized that insights into the workings of complex systems like perception, memory, and language require an understanding of the general operating principles of the system (Anderson, 1990; Chomsky, 1995; Marr, 1982). Probabilistic models, which use tools from Bayesian statistics and machine learning to describe such systems, represent a promising framework for exploring high-level descriptions of cognitive processes (Chater et al., 2006; Tenenbaum et al., 2011).[1]

Although probabilistic models have grown tremendously in popularity in recent years, they have also attracted significant criticism (Bowers and Davis, 2012; Jones and Love, 2011). Chief among these criticisms is that these models imply a claim that the mind itself is *rational* or even

[1] I use the terms "probabilistic" and "Bayesian" synonymously. I prefer "probabilistic," as it better describes the key virtue of these models: that they use probability as a single framework for integrating across widely varying tasks, representations, and constraints.

* Tel.: +1 650 724 4003.
E-mail address: mcfrank@stanford.edu

0010-0277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2013.04.010

optimal. A claim of optimality entails that a particular cognitive process provides the best possible solution relative to some problem. The weaker claim of rationality suggests that the process provides a logical, well-designed solution to a problem, perhaps relative to limitations on cognitive resources like memory or computation. These claims—especially the optimality claim—strike many authors as unsupported and unfalsifiable, given that the particular problem being solved and the assumptions of the model solving it are rarely specified independently. Endress's (2013) article echoes these criticisms of optimality claims, applying them to Frank and Tenenbaum's (2011) models of sequential rule learning (henceforth, "FT") and providing additional theoretical and empirical arguments.

2. "Rule learning" and Endress's critique

FT used probabilistic models to describe infants' and adults' ability to learn sequential regularities in auditory stimuli, a learning ability that may be linked to language acquisition (Marcus et al., 2007; Marcus et al., 1999; Peña et al., 2002). In "rule learning" paradigms (Marcus et al., 1999), learners hear strings of syllables like "wo fe fe," instantiating simple regularities (e.g., in this case *ABB*, or "last syllable repeats"). They are then tested on their ability to generalize these regularities to novel syllable strings. Experiments across a variety of ages, modalities, and rule types provide a rich body of data that can be explored for insights about how infants and adults make such generalizations (Endress et al., 2005; Gerken, 2006; Johnson et al., 2009; Marcus et al., 2007; Marcus et al., 1999).

FT created three probabilistic models that made predictions about learners' performance across a wide range of empirical results. All three of these models were based on the assumption that learners prefer more specific hypotheses (the "size principle" of Tenenbaum and Griffiths, 2001), but they varied in their complexity. Model 1, the simplest, made inferences directly from the input data, but it always learned the correct rule perfectly. Model 2 added a single free parameter that controlled noise in memory, allowing the model to produce quantitative predictions. Model 3 learned multiple rules. Despite the apparent diversity of results in the literature, these simple models sufficed to describe a wide range of empirical data. Our explicit goal in FT was to provide "a baseline for future work that can be modified and enriched" as the data warranted.

Endress (2013) argues that our models do not provide a good account of existing data on rule learning, however, contesting both the general framework we used and the specifics of our simulations. In this response I will focus primarily on a set of critiques that have broad interest:

1. Learners prefer more salient rules rather than more specific hypotheses.
2. "Common-sense psychology" provides an adequate explanation of rule learning.
3. The use of free parameters is inappropriate in cognitive modeling.
4. The use of probabilistic models implies an optimality claim.

In Appendix A, I briefly summarize responses to criticisms of specific simulations.

To summarize, I argue that Endress's criticisms 1–3 are not valid. Moving beyond a notion of "common sense" psychology to theories that make graded and quantitative predictions, we will need to use statistical tools to understand and evaluate the flexibility and specificity of our theories. Nevertheless, Endress's article raises useful questions about how computational principles can be instantiated in human minds and I am in agreement that there should be a high standard for claims of optimality on the basis of probabilistic modeling (indeed, FT did not make such a claim).[2]

3. Do learners prefer more specific rules?

At the heart of Endress's critique is the claim that "humans do not prefer more specific patterns." This claim is important because the size principle (Tenenbaum and Griffiths, 2001; Xu and Tenenbaum, 2007b)—the principle that hypotheses are weighted proportional to their specificity, as a consequence of how those examples are sampled—was the major explanatory assumption in FT's models.

A large, independent body of evidence supports the use of the size principle as a description of word learning and categorization (Navarro et al., 2012; Tenenbaum and Griffiths, 2001; Xu and Tenenbaum, 2007a, 2007b) and the sensitivity of even young infants to the sampling processes that result in the size principle (Denison et al., 2012; Gweon et al., 2010; Kushnir et al., 2010; Xu and Garcia, 2008; Xu and Denison, 2009). To take just one example, in the word learning tasks used by Xu and Tenenbaum (2007b), adults and children saw either one or three examples of a category and were asked to make judgements that revealed the specificity of their generalization. Presented with one example, they showed gradient generalization, but after seeing three examples, their judgments were consistent with the most specific category that fit the data they observed. This dataset and many others provide powerful evidence for the importance of strong sampling and the size principle, but are not discussed by Endress.

Instead, in support of the claim that humans do not prefer more specific rules, Endress conducted an experiment in which participants were familiarized with human speech syllables in an *AAB* or *ABB* pattern. At test they were asked to choose between pattern-incongruent human syllables, or pattern-congruent strings instantiated in rhesus monkey vocalizations, pitting consistency with the pattern regularity (e.g., *AAB* vs. *ABB*) against consistency in the modality of presentation (human speech vs. monkey

[2] This response represents my personal views. There is substantial variance in attitudes towards optimality claims in the probabilistic modeling literature, and the attitudes of many researchers have evolved with respect to this issue as more research has focused on "process level" explanations for cognitive phenomena (Chater et al., 2011; Griffiths et al., 2012; Sanborn et al., 2010; Vul et al., 2009).