How are Bayesian models really used? Reply to Frank (2013)

Ansgar D. Endress
Universitat Pompeu Fabra, Barcelona, Spain
City University, London, UK
Massachusetts Institute of Technology, Cambridge, MA

Draft of September 12, 2013. Please do not quote without permission.

In response to the proposal that cognitive phenomena might be best understood in terms of cognitive theories (Endress, 2013), Frank (2013) outlined an important research program, suggesting that Bayesian models should be used as rigorous, mathematically attractive implementations of psychological theories. This research program is important and promising. However, I show that it is not followed in practice. I then turn to Frank's defense of the assumption that learners prefer more specific rules (the "size principle"), and show that the results allegedly supporting this assumption do not provide any support for it. Further, I demonstrate that, in contrast to Frank's criticisms, there is no circularity in an account of rule-learning based on "common-sense psychology", and that Frank's other criticisms of this account are unsupported. I conclude that the research program outlined by Frank is important and promising, but needs to be followed in practice. Be that as it might, the rule-learning experiments discussed by Frank are still better explained by simple psychological mechanisms.

Introduction

Using Frank and Tenenbaum's (2011) (hereafter FT) Bayesian model of rule learning as a case study, I (Endress, 2013) proposed that such models cannot be considered ideal-observer models but rather make important ad-hoc assumptions that determine the model behavior. Further, I argued that a model based on simple psychological mechanisms explains the data better. Frank (2013) outlined a general strategy for how Bayesian models could make important contributions for studying cognition. While promising and important, I show that Frank's research program is generally not followed in practice, and that Frank's criticisms of my other points and of the simple psychological model of rule learning are unsupported by earlier research and, therefore, unfounded (see Appendix A for specific replies to Frank's claims; a critical discussion of the evidence Frank cites in support of the size principle can be found in Endress, under review).

How to use Bayesian models for studying cognition

Frank proposes that Bayesian models are well suited for formalizing theories of cognition, by implementing hypotheses in a framework with attractive mathematical properties, as "Bayesian inference is 'optimal' in the sense that it leads to the correct posterior distribution." In other words, using a Bayesian methodology guarantees that, given a set of assumptions, the predictions of Bayesian models are indeed those that follow from the assumptions. This is a useful property, even though non-Bayesian models make the predictions that follow from their assumptions as well. Crucially, however, while this approach is promising and important, I will show below that it is rarely followed in practice, for two important reasons.

First, if Bayesian models were really used as suggested by Frank, they would be silent on issues about whether human information processing is optimal. While Frank argues that Bayesian models are not used to make such claims of optimality, he also acknowledges that many modelers do make such claims. In fact, even FT assert that their models reflect the "computational structure of the task" (Footnote 1); if so, learners who behave according to normatively correct inferences based on the computational structure of the task presumably behave "optimally" as well. Hence, in many cases, Bayesian models are not used as ideal-observer models, but rather to draw conclusions that are, according to Frank, rarely justified.

The second reason relates to the goal of Bayesian

This research was supported by NIH grant MH47432, grants CONSOLIDER-INGENIO-CDS-2007-00012 and PSI2012-32533 from the Spanish Ministerio de Economía y Competitividad, SGR-2009-1521 from the Catalan government, and Marie Curie Incoming Fellowship 303163-COMINTENT. I am grateful to M. Frank for helpful comments on an earlier draft of this manuscript.

models. FT (and many other Bayesian modelers) propose their models to be higher-level models of the computational structure of the problem that are agnostic about the underlying mechanisms. However, FT continuously switch between levels of description, making it difficult to decide what the model actually describes. For example, while presenting their model as an ideal-observer model, FT also hold that infants are batch learners who remember all familiarization items before making inferences.[1] This requires them to include further implementational parameters about memory reliability, and these parameters have a substantial effect on the model behavior in turn. Hence, by continuously switching between levels of description, it becomes extremely difficult to decide whether the model behavior is due to the theory FT set out to test, or rather to one of the extraneous assumptions.

This is particularly clear in FT's use of the forgetting rates in their simulations. For example, for theorists who, like FT, hold that learners are guided by the size principle and faithfully remember familiarization items, it might be important to find out under which conditions the models account for the data, and, in fact, the models provide rich information about this issue. In some simulations, FT need to assume forgetting rates of 10%, in others of 40%, in others of 60% and in still others of 80%, and the models often do not fit the data unless such specific forgetting rates are assumed. If, as Frank suggests, "investigating the dependence of predictions on assumptions about perceptual and memory noise is precisely the purpose of ideal observers," a plausible conclusion from the models' fatal dependence on very specific and mutually inconsistent parameter values is that the models do not in fact provide adequate accounts of the experimental data.

Is there evidence that learners prefer more specific rules?

One of the crucial assumptions of FT's models, and one of the crucial arguments of Frank's reply, is that learners preferentially learn more specific rules (the "size principle"). As I pointed out, the size principle has sound justifications in the case of language acquisition (e.g., Hyams, 1986; Manzini & Wexler, 1987), but its use by FT is particularly implausible. Specifically, in FT's model, infants might encounter a total of three syllables. Before encountering any syllable triplet, infants know that the three syllables allow for a total 27 triplets, that 6 of these triplets follow an ABB pattern (e.g., *pu-li-li*), that 3 of these triplets follow an AAA pattern (where all three syllables are identical), as well as the number of triplets that would conform to any conceivable rule. Unless infants have innate knowledge of the number of items compatible with any conceivable rule and any conceivable number of syllables, FT's models suggest that infants have to process about 900 hypothetical and counterfactual triplets per second. This assumption lies at the core of FT's model. Hence, if it is unsupported, the model become unsupported as well.[2]

While Frank claims that several papers support the size principle, most of the data presented in these papers is unrelated to the size principle, confounded by other factors, can be fit by more plausible models that make no use of the size principle, or is inconsistent with the models it is allegedly predicted by. A critical discussion of these papers can be found elsewhere (Endress, under review). Further, even if these results supported some version of the size principle, they would not support FT's models, because participants could estimate the size of classes based on stimuli they actually saw, and not on hundreds of thousands of hypothetical and counterfactual stimuli they never experienced.

Given the absence of evidence for the size principle as used by FT, I ran an experiment illustrating the fact that the size principle cannot be taken for granted. Participants were presented with a sequence of syllable triplets conforming to a repetition-pattern (e.g., *wo-fe-fe*). Following this, they had to choose between triplets of rhesus monkey vocalizations conforming to the pattern, and triplets of human syllables violating the pattern. As a result, they could choose between the more specific repetition-pattern, and the less specific "all items are made of syllables" regularity. Importantly, FT's incorporated both types of rules; hence, this experiment compared the predictions of FT's model to the behavior of actual participants. However, results showed that most participants chose the less specific "all

[1] Indeed, FT "distinguished the larger memory demands involved in maintaining a representation of training items across a long exposure period compared with an individual evaluating test items in the moment". However, if the memory demands are different in different parts of the experiment, memory is necessarily used in the process of the generalizations. As a result, infants must be batch learners who remember all familiarization items (more or less faithfully according to the memory parameter).

[2] While FT acknowledged the psychological implausibility of such a model, Frank argues that enumerating all possible triplets is not implausible after all, because "research on numerical cognition suggests that adults and infants need not enumerate to make quick and accurate judgments about the cardinality of sets (Xu, 2002; Whalen, Gallistel, & Gelman, 1999)." However, participants in number cognition experiments have to process maybe up to 100 dots on a screen, but usually much fewer, and certainly not hundreds of thousands of items as in FT's simulations. Further, and crucially, observers in numerical cognition tasks *see* the objects they have to enumerate; in contrast, in FT's simulations, the triplets to be enumerated are hypothetical and counterfactual. It seems fair to assume that observer cannot estimate the number of dots on displays they are never shown. Likewise, there is no evidence that participants can "count" the number of triplets they never hear. As a result, the number processing literature provides no support for FT's models.

items are made of syllables" regularity.

Frank dismisses these data, arguing that they "do not provide evidence against the size principle." However, this was not the point of the experiments. Rather, I argued that they "fail to support the predictions of Frank and Tenenbaum's (2011) model and demonstrate that a preference for more specific patterns cannot be taken for granted," and that the size principle can be easily overwritten by other stimulus properties. Hence, Frank's claims notwithstanding, there is no evidence for the central assumption of FT's models, which, therefore, remains speculative and psychologically implausible.

Common sense psychology does account for the data

An alternative approach to rule-learning can be based on perceptual or memory primitives (Endress, Scholl, & Mehler, 2005; Endress, Dehaene-Lambertz, & Mehler, 2007; Endress, Nespor, & Mehler, 2009). Frank criticizes this account for being circular, because it does not specify which patterns learners might pick up. He gives the example that, when presented with syllable triplets conforming to a repetition-pattern, participants might well learn generalizations "like 'any string that ends in /di/, /je/, /li/, or /we/' or 'any string with three or four elements.' "

However, the circularity results from a confusion between the patterns that are learned, and those that are tested. In fact, to the extent that infants can detect the number of sounds (or properties with which the number of sounds might be confounded; see Starkey, Spelke, & Gelman, 1983, 1990, but see Lipton & Spelke, 2004), they will do so also in rule-learning experiments. Likewise, Gerken's (2010) data clearly show that infants can learn a regularity of the sort "any string that ends in /di/", and adults can, at least after many more familiarization examples, learn regularities of the sort "any string that ends in /di/, /je/, /li/, or /we/" as well (Endress & Mehler, 2009). However, such abilities were simply not tested in these experiments.

In fact, the psychological account does not need to make arbitrary assumptions. The question of which hypotheses infants can and do entertain is simply an empirical one. Likewise, infants' behavior will plausibly be mostly driven by the most salient patterns; the relative saliency of different pattern is an empirical question, and can be tested by pitting patterns against each other (see Gervain & Endress, in preparation). Of course, such an approach does not answer the question of *why* infants entertain some hypotheses, or why some patterns are more salient than others. However, while FT propose an answer to such questions, their explanation either makes incorrect predictions, or does not account for the data to be explained in the first place.

Conclusions

Frank proposes an important step forward in the use of Bayesian models to study cognition, namely to use them to refine hypotheses and make testable predictions. If this were how Bayesian models are used in practice, Franks proposal could become an important tool for developing psychological theories of psychological phenomena and to ground them in empirical research.

Appendix A
Responses to specific responses

In his Appendix A, Frank comments on some of my more specific criticisms of FT's models. I will briefly comment on his responses in turn. As pointed out by Frank, many of the criticism were related to the lack of evidence for the size principle, and the associated implausibility of the model. As discussed above, there still is no such evidence.

Regarding Endress et al.'s (2007) data, FT reproduced the performance difference between repetition-patterns and "ordinal" patterns, arguing that it might be due to the fact that, in the case of the ordinal patterns, "a number of possible rules were consistent with the training stimuli" (p. 365). If so, one would expect the same difficulties with simple rising or falling contours such as "lowest-middle-highest." I showed that this is not the case, and that most participants are at ceiling learning rising or falling contours. However, Frank's new simulations show that raising and falling contours are learned better than ordinal patterns. Still, further simulations revealed that, if, as in FT, one treats the relative surprisal for incorrect vs. correct test item as a measure of performance on the test items and uses FT's model parameters, the model predicts that performance for repetition-patterns should be about 20% better than for rising or falling contours for the number of tones used by Endress et al. (2007), and can be made arbitrarily large simply by changing the number of tones from which the triplets are constructed.[3] Given that most participants were at ceiling with the rising and falling contours, the data contradict this prediction. Hence, the conclusion is inevitable that FT's models reproduced Endress et al.'s (2007) results due to assumptions that are empirically incorrect.

Regarding Gerken's (2010) data, Frank acknowledges that the model's ability to learn from a few examples results in the prediction that humans should unlearn a rule from a single counter-example, even after thousands of positive examples. While Frank argues that this is a short-coming of the models that can be fixed,

[3] For example, with 18 tones, participants should perform about 42% better on the repetition patterns than on the rising and falling contours. This model behavior is expected, as it is easy to calculate that the rising/falling regularity is less specific than the repetition pattern for 9 tones or more.

this "short-coming" is actually the reason for which the model accounted for the data in the first place.

Regarding Gómez's (2002) results, Frank acknowledges that the model fails to account for the data, but argues that this can be addressed in further work. However, given that, according to Frank, this revolves around the interpretation of one of their parameters, it is unclear why FT introduced a parameter with no clear interpretation in the first place.

Regarding Kovács and Mehler's (2009a) data, Frank argues that ideal-observer are not opposed to process models, even though FT's ideal observer model makes no contact whatsoever with well-established process-based explanation. Specifically, Kovács and Mehler (2009b) showed that bilingual infants can learn multiple regularities while monolinguals can learn only one, an ability that has been linked to the better developed executive function in bilinguals. FT modified their model to be more likely to admit more than one regularity, found that it was indeed more likely to learn more than one regularity, and conclude that bilinguals are somehow designed to be more likely to admit more than one regularity as well. Further, they postulate that bilingual advantages in tasks such as the Stroop task are a consequence of being more likely to learn multiple regularities as well, with no supporting evidence whatsoever, despite the prima facie implausibility of this claim.

Appendix B
References

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159–176.

Endress, A. D. (under review). Is there evidence for the size principle? a critical review.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

Endress, A. D., & Mehler, J. (2009). Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology*, *62*(11), 2187–2209.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348–353.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology. General*, *134*(3), 406-19.

Frank, M. C. (2013). Throwing out the bayesian baby with the optimal bathwater: Response to. *Cognition*, *128*(3), 417–423.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371.

Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, *115*(2), 362-6.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431-6.

Hyams, N. (1986). *Language acquisition and the theory of parameters.* Dordrecht: D. Reidel.

Kovács, A. M., & Mehler, J. (2009a). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6556–6560.

Kovács, A. M., & Mehler, J. (2009b). Flexible learning of multiple speech structures in bilingual infants. *Science*, *325*(5940), 611–612.

Lipton, J. S., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, *5*(3), 271–290.

Manzini, M. R., & Wexler, K. (1987). Parameters, binding theory, and learnability. *Linguistic Inquiry*, *18*(3), pp. 413-444.

Starkey, P., Spelke, E., & Gelman, R. (1983). Detection of intermodal numerical correspondences by human infants. *Science*, *222*(4620), 179-81.

Starkey, P., Spelke, E., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition*, *36*(2), 97-127.

Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*(2), 130-137.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223–250.