**How robust are probabilistic models of higher-level cognition?**

6 January 2013

Gary F. Marcus

Ernest Davis

New York University

Abstract: *An increasingly popular theory holds that the mind should be viewed as a “near optimal” or “rational” engine of probabilistic inference, in domains as diverse as categorization, word learning, pragmatics, naïve physics, and predictions of the future. We argue that this view, often identified with Bayesian models of inference, is markedly less promising than widely believed, undermined by post hoc practices that merit wholesale reevaluation. We also show the common equation between* *probabilistic and “rational” or “optimal” is not justified.*

Should the human mind be seen as an engine of probabilistic inference, yielding “optimal” or “near-optimal” performance, as several recent, prominent articles have suggested (Frank & Goodman, 2012; Gopnik, 2012; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Téglás et al., 2011)? Tenenbaum et al (2011) argue that

*Over the past decade, many aspects of higher-level cognition have been illuminated by the mathematics of Bayesian statistics: our sense of similarity (18), representativeness (19), and randomness (20); coincidences as a cue to hidden causes (21); judgments of causal strength (22) and evidential support (23); diagnostic and conditional reasoning (24, 25); and predictions about the future of everyday events (26)… [as well as] perception (27), language (28), memory (29, 30), and sensorimotor systems (31)* [references in original]

In support of this view, experimental data have been combined with precise, elegant models that provide remarkably good quantitative fits. For example, Xu and Tenenbaum (Xu & Tenenbaum, 2007) presented a well-motivated probabilistic model “based on principles of rational statistical inference” that closely fits adult and children’s generalization of novel words to categories at different levels of abstraction (*green pepper* vs *pepper* vs *vegetable)*, as a function of how labeled examples of those categories are distributed.

In these models, cognition is viewed as a process of drawing inferences from observed data in a fashion normatively justified by mathematical probability theory. In probability theory, this kind of inference is governed by Bayes’ law. Let D be the data and H1 … Hk be hypotheses; assume that it is known that exactly one of the Hi is true. Bayes’ rule states that, for each hypothesis Hi,



In this equation P (Hi|D) is the *posterior* probability of the hypothesis Hi given that the data D have been observed. P(Hi) is the *prior* probability that Hi is true before any data have been observed. P(D| Hi) is the *likelihood*; the conditional probability that D would be observed assuming that Hi is true. The formula states that the posterior probability is proportional to the product of the prior probability times the conditional probability. In most theories discussed here: the "data'' are information available to a human reasoner, the "priors" are a characterization of the reasoner's initial state of knowledge and the "hypotheses'' are the conclusions that he/she draws. For example, in word learning task, the data could be observations of language and a hypothesis could be a conclusion that the word "dog" denotes a particular category of object (friendly, furry animals that bark.)

Couching their theory in the language of evolution and adaptation, Tenenbaum et al argue (2011) argue that

*The Bayesian approach [offers] a framework for understanding why the mind works the way it does, in terms of rational inference adapted to the structure of real-world environments.*

To date, these models have been criticized only rarely (Jones and Love, 2011; Bowers and Davis, 2012; Eberhardt and Danks, 2011). Here, through a series of detailed case studies, we demonstrate that two closely related problems - one of task selection, the other of model selection -- undermine any general conclusions about whether cognition is in fact either optimal or driven by probabilistic inference. Furthermore, we show that multiple probabilistic models are often potentially applicable to any given task (some compatible with the observed data but others not), that published claims of fits of probabilistic models sometimes depend on post hoc choices that are unprincipled, and that, in many cases, extant models depend on assumptions that are empirically false, nonoptimal, or both.

**Task Selection**

In a recent study of physical reasoning, Hamrick et al (Hamrick, Battaglia, & Tenenbaum, 2011) asked subjects to assess the stability of towers of blocks. Participants were shown a computer display showing a randomly generated three-dimensional tower of blocks and asked to predict whether it was stable or would fall, and, if it fell, in what direction it would fall.

Hamrick et al. proposed a model according to which human subjects correctly use and represent Newtonian physics, with errors arising only to the extent that subjects are affected by perceptual noise, in which the perceived x and y coordinates of a block vary around the actual position according to a Gaussian distribution.

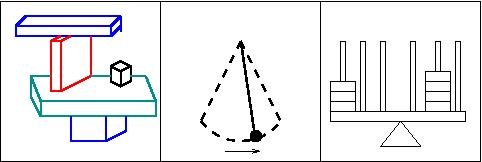


Figure 1: Three tests of intuitive physics. Panel A: Estimating the stability of towers of blocks. Panel B: Estimating the trajectory of projectiles. Panel C:Estimating balance. Human subjects do well in scenario A, but not B or C.

Within the set of problems studied, the model closely predicts human data, and the authors conclude that “Though preliminary, this work supports the hypothesis that knowledge of Newtonian principles and probabilistic representations are *generally* applied for human physical reasoning" [emphasis added].

The trouble with such claims is that human cognition often seems near-normative in some circumstances but not others. A substantial literature, for example, has already documented human difficulties with respect to other Newtonian problems (McCloskey, 1983). For example, one study (Caramazza, McCloskey, & Green, 1981) asked subjects to predict what would happen if someone were spinning a rock on a string, and then released the string. The subjects mostly predicted that the rock would then follow a circular or spiral path, rather than the correct answer that the trajectory of the rock would be the tangent line. Taken literally, Hamrick's claim would predict that subjects should be able to answer this problem correctly; it also overestimates subjects’ ability to predict accurately the behavior of gyroscopes, coupled pendulums, and cometary orbits.

As a less challenging test of the generalizability of the Hamrick et al. probabilistic-Newtonian approach, we applied the Hamrick et al. model to balance beam problems (Figure 1C). These involve exactly the same physical principles; therefore, Hamrick et al.’s theory should predict that any subject errors can be accounted for in terms of perceptual uncertainty. We applied Hamrick et al.’s model (Gaussian distribution) of uncertainty to positional and mass information, both separately and combined. The result was that, for a wide range of configurations, given any reasonable measure of uncertainty (see supplement), the model predicts that subjects will always predict the behavior correctly.

As is well-known in the experimental literature, however, this prediction is false. Both children and many untutored adults (Siegler, 1976) frequently make a range of errors, such as relying solely on the number of weights to the exclusion of information about how far those weights are from the fulcrum. On this problem, only slightly different from that posed by Hamrick (both hinge on factors about weight, distance, and leverage), the fit of Hamrick’s model is very poor. What held true in the specific case of their tower problems -- that human performance is near optimal -- simply is not true in a problem governed by the laws of physics applied in a slightly different configuration. (Of course sophisticated subjects, such as Hamrick’s et al’s pool of MIT-trained undergraduates may do better.)

The larger concern is that the probabilistic cognition literature as a whole may disproportionately report successes, akin to Rosenthal’s file drawer problem (Rosenthal, 1979), leading to a distorted perception of the applicability of the approach. Table 1 accumulates many of the most influential findings in the cognitive literature on probabilistic inference, and shows that, in the vast majority, results that fit naturally with probabilistic techniques and claims of optimality are closely paralleled with other equally compelling results that do not fit so squarely, raising important issues about the generalizability of the framework.

**Table 1:** *Examples of phenomena in different domains that do and do not fit naturally with probabilistic explanations*

|  |  |  |
| --- | --- | --- |
| **Domain** | **Apparently optimal** | **Apparently non-optimal** |
| Intuitive physics | towers (Hamrick et al., 2011) | balance-scale (Siegler, 1976)  projectile trajectories (Caramazza et al., 1981) |
| Incorporation of base rates | various (Frank & Goodman, 2012; Griffiths & Tenenbaum, 2006); | base rate neglect (Kahneman & Tversky, 1973) [but see (Gigerenzer & Hoffrage, 1995)] |
| Extrapolation from small samples | future prediction (Griffiths & Tenenbaum, 2006) size principle (Tenenbaum & Griffiths, 2001a) | anchoring (Tversky & Kahneman, 1974)  underfitting of exponentials (Timmers & Wagenaar, 1977) gambler's fallacy  conjunction fallacy (Tversky & Kahneman, 1983)  estimates of unique events (Khemlani, Lotstein, & Johnson-Laird, 2012) |
| Word learning | sample diversity (Xu & Tenenbaum, 2007) | sample diversity (Gutheil & Gelman, 1997)  evidence selection (Ramarajan, Vohnoutka, Kalish, & Rhodes, 2012) |
| Social cognition | pragmatic reasoning (Frank & Goodman, 2012) | attributional biases (Ross, 1977)  egocentrism (Leary & Forsyth, 1987)  behavioral prediction (children) (Boseovski & Lee, 2006) |
| Memory | rational analysis (Anderson & Schooler, 1991) | eyewitness testimony (Loftus, 1996)  vulnerability to interference (Wickens, Born, & Allen, 1963) |
| Foraging | animal behavior (McNamara, Green, & Olsson, 2006)  “information-foraging” (Jacobs & Kruschke, 2011) | probability matching (West & Stanovich, 2003) |
| Deductive reasoning | deduction (Oaksford & Chater, 2009) | deduction (Evans, 1989) |
| Overview | higher-level cognition (Tenenbaum et al., 2011) | higher-level cognition (Kahneman, 2003; Marcus, 2008) |

The risk of confirmationism is almost certainly exacerbated by the tendency of advocates of probabilistic theories of cognition (like researchers in many computational frameworks) to follow a breadth-first search strategy -- in which the formalism is extended to an ever-broader range of domains (most recently, intuitive physics and intuitive psychology) -- rather than a depth-first strategy in which some challenging domain is explored in great detail with respect to a wide range of tasks.

More revealing than picking out arbitrary tasks in new domains might be deeper exploration of domains that juxtapose large bodies of “pro” and “anti” rationality literature. For example, when people extrapolate, they are sometimes remarkably accurate, as Griffiths and Tenenbaum (Griffiths & Tenenbaum, 2006) have shown, but at other times remarkably inaccurate, as when they “anchor” their judgments based on arbitrary and irrelevant bits of information. (Tversky & Kahneman, 1974) An attempt to understand the seemingly competing mechanisms involved might be more illuminating than the current practice of identifying a small number of tasks in each domain that seem to be compatible with a probabilistic model.

**Model Selection**

Closely aligned with the problem of how tasks are selected is the problem of how *models* are selected. Each model depends heavily on the choice of probabilities; these probabilities can come from three kinds of sources:

1. Real-world frequencies
2. Experimental subjects' judgments
3. Mathematical models, such as Gaussians or information theoretic arguments

A number of other parameters must also be set either by basing the model/parameter on real world statistics either for this problem or for some analogous problem; by basing the model/parameter on some other psychological experiment; by choosing the model/tuning the parameter to best fit the experiment at hand; or by using purely theoretical considerations, sometimes quite arbitrary.

Unfortunately, each of these choices can be problematic. To take one example, real world frequencies may depend very strongly on the particular data set being used or the sampling technique or the implicit independence assumptions. For instance, Griffiths and Tenenbaum (Griffiths & Tenenbaum, 2006) studied estimation abilities. Subjects were asked questions like “If you heard that a member of the House of Representatives had served for 15 years, what would you predict his total term in the House would be?” Correspondingly, a model was proposed in which the hypotheses are the different possible total length of term; the prior is the real-world distribution of the lengths of representatives' terms; and the datum is the fact that the representative's term of service is at least 15 years; and analogously for the other questions. In seven of nine questions, these models accounted very accurately for the subjects' responses. Griffiths and Tenenbaum concluded that “everyday cognitive judgments follow the … optimal statistical principles … [with] close correspondence between people’s implicit probabilistic models and the statistics of the world. ”

But it is important to realize that the fit of the model to the data depend heavily on how the priors are chosen. To the extent that priors may be chosen post hoc, the true fit of a model can easily be overestimated, perhaps greatly. For instance, one of the questions in the study was "If your friend read you her favorite line of poetry and told you it was line 5 of a poem, what would you predict for the total length of the poem?" How well the model fits the data depends on what prior is presupposed. Griffiths and Tenenbaum based their prior on the distribution of length on an online corpus of poetry (http://www.emule.com). To this distribution, they applied the following stochastic model,

motivated by Tenenbaum’s "size principle": It is assumed, first that the choice of "favorite line of poetry" is uniformly distributed over poems in the corpus, and second that, given the poem, the choice of favorite line is uniformly distributed over the lines in the poem. Finally, it is assumed the subjects' answer to the question will be the median of the posterior distribution.

From the apparent fit, Griffiths and Tenenbaum claim that "People's judgements for ... poem lengths ... were indistinguishable from optimal Bayesian predictions based on the empirical prior distributions." However, the fit between the model and the experimental results is not in fact as close as the diagram in Griffiths and Tenenbaum would suggest. (They did not do a statistical analysis, just displayed a diagram.) In their diagram of the results of this experiment, the y-axis represents the total length of the poem, which is the question that the subjects were asked. However, it requires no great knowledge of poetry to predict that a poem whose fifth line has been quoted must have at least five lines; nor will an insurance company pay much to an actuary for predicting that a man who is currently thirty-six years old will live to at least age thirty-six. The *predictive* part of these tasks is to predict how much longer the poem will continue, or how much longer the man will live. If the remaining length of the poem is used as the y-axis, as in the right-hand panel in figure 2, it can be seen that though the model has some predictive value for the data, the data is by no means "indistinguishable" from the predictions of the model.

More importantly, the second assumption in the stochastic model above, that favorite lines are uniformly distributed throughout the length of a poem, is demonstrably false. Examining an online data set of favorite passages of poetry (http://www.poets.org/page.php/prmID/339), it is immediately evident that favorite passages are *not* uniformly distributed; rather, favorite passages are generally the first or last lines of the poem, with last lines being about twice as frequent as first lines. As illustrated in the right hand panel of Figure 2, a model that incorporated these empirical facts would yield a very different set of predictions. Without independent data on subject’s priors, it is impossible to tell whether the Bayesian approach yields a good or a bad model, because the model’s ultimate fit depends entirely on which prior subjects might actually represent.

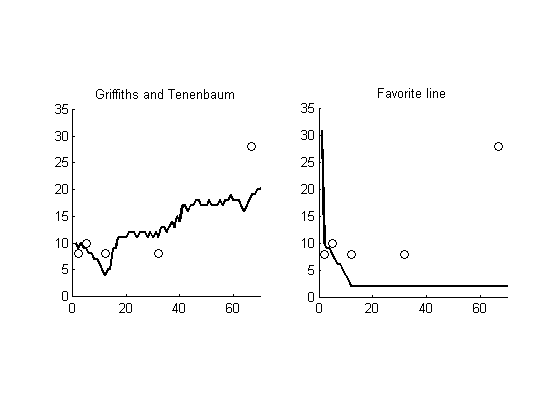


Figure 2: Two different predictions the probabilistic model could make, depending on how priors were selected. The solid line in the figure on the left shows the predictions made by the model in Griffiths and Tenenbaum. The solid line in the figure on the right shows predictions made based on empirical data about distributions of favorite lines. In both figures, the small figures correspond to the mean of the subjects' responses. The y-axes are the models’ predictions of the number of lines that remain in the poem *after* the chosen line, not the total number of lines in the poem.

The analysis of movie gross earnings is likewise flawed. Subjects were asked, “Imagine you hear about a movie that has taken in 10 million dollars at the box office, but don’t know how long it has been running. What would you predict for the total amount of box office intake for that movie?” The data set used was a record of the gross earnings of different movies. The fit of the probabilistic model is conditioned on the assumption that movie earnings are uniformly distributed over time; for example, if the film earns a total of $100 million, then the question about this movie is equally likely to be raised after it has earned $5 million, $10 million, $15 million and so on. But movies, particularly blockbusters, are heavily front-loaded and earn most of their gross during the beginning of their run. No one ever heard that *The Dark Knight* (total gross $533M) had earned $10 million, because its gross after the first three days was $158 million. Factoring this in would have lead to a different prior (one in which projected earnings would be substantially lower) and a different conclusion (one in which subjects overestimated future movie earnings, and hence were *not* optimal). To put this another way: The posterior distribution used by Griffiths and Tenenbaum corresponds to a process in which the questioner first picks a movie at random, then picks a number between zero and the total gross, and then formulates his question. However, if instead, the questioner randomly picks a movie currently playing and formulates his question in terms of the amount of money it has earned so far, then the posterior distribution of the total gross would be very different, since, due to the front-loading, most of the movies playing at any given moment have earned most of their final gross. Again, we cannot legitimately infer that the model is accurate without independent evidence as to a subject’s priors.

Seemingly innocuous design choices can yield models with arbitrarily different predictions in other ways, as well. Consider for instance a recent study of pragmatic reasoning and communication (Frank & Goodman, 2012) which purported to show that “speakers act rationally according to Bayesian decision theory.” Subjects were shown sets of three objects, and asked to place certain bets pertaining to the likelihood that certain words being used in particular ways, e.g., would a speaker use the word “blue” to pick out the middle object in the tableau depicted in the left panel of Figure 3?

Imagine someone is talking to you and uses [the word “blue”/a word you don’t know]   
to refer to one of these objects. Which object are they talking about?

Figure 3: From (Frank and Goodman, 2012)

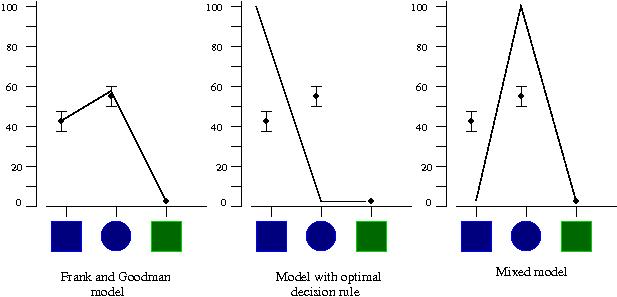


Figure 4: Analysis of the effect of varying decision rules. Left panel: Predictions given proportional decision rule, used in (Frank and Goodman, 2012) that is suboptimal. Center panel: Predictions given that the hearer assumes that the speaker follows the winner-take-all (MEU) rule. (Suppose the speaker follows the MEU rule. Then, if the center object is chosen, the speaker will certainly say “circle” rather than “blue”. Hence, if the speaker says “blue” he must be referring to the object on the right.) Right panel: Prediction given assuming that the hearer assumes that the speaker will follow a proportional decision rule, but the hearer follows an MEU rule. As shown, the subject’s response is not actually optimal (contra claims made in Frank and Goodman), and the fit of the model varies considerably depending on the post hoc choice of decision rule. The y-axis here represents the amount of money bet on the prediction; e.g. in the left panel, the hearer bets $43 that the object the speaker is referring to is the blue square, $57 that it is the blue circle, and $0 that it is the green square.

Frank and Goodman showed that a probabilistic “rational actor" model of the speaker, with utility defined in terms of “surprisal” (a measure of the information gained by the hearer) could predict subject performance with near perfect accuracy (Figure 4, left panel). The trouble is that their model of these data depends critically on a “decision rule” that speakers choose the word to use with a probability proportional to the word’s specificity; In this particular example, "blue" has a specificity of 1/2, since it applies to two objects, and "circle" has a specificity of 1, since it applies to only one; therefore speakers who wish to specify the middle object would use "circle" 2/3 of the time and "blue" 1/3 of the time. Although this decision rule is not uncommon, Frank and Goodman might have just as easily have chosen a “winner-take-all” decision rule in which subjects supplied only the most likely response; in this example "circle". The "winner-take-all" rule follows the maximum expected utility (MEU) principle, which is the standard rule in decision theory. As Figure 4 shows, echoing Figure 3, although the model with the published decision rule yields a good fit to the data, other models, actually more justifiable a priori (see Figure caption), would have yielded dramatically poorer fits.

The experimenters can also manipulate the choice of wording. For example, rather than asking the subject which word they would be most likely to use (which seems ecologically natural) Frank and Goodman (M. C. Frank, pers comm, 21 Sept 2012) asked subjects to "Imagine that you have $100. You should divide your money between the possible words -- the amount of money you bet on each option should correspond to how likely you would be to use that word. Bets must sum to 100!” In effect this asks the subject to place a bet on what they themselves would say, ecologically anomalous and coercive, in that the phrasing "should divide" places a task demand such that all-or-none-answers are pragmatically discouraged. Had the question instead been posed as *”Imagine that you are talking to someone and you want to refer to the middle object. Which word would you use, “blue” or “circle?”*, we suspect that 100% of the speakers would have answered “circle” rather than the 67% observed. (Saying “blue” would be a violation of Gricean constraints, and an actual hearer would complain that this was ambiguous or misleading.)

Table 2 enumerates some of the features that have empirically varied (without strong a priori theoretical basis) across probabilistic models of human cognition. Individual researchers are free to tinker, but the collective enterprise suffers if choices across domains and tasks are unprincipled and inconsistent. Models that have been fit only to one particular set of data have little value if their assumptions cannot be verified independently; otherwise, the entire framework risks becoming an exercise in squeezing round pegs into square holes.

(Another vivid example of a Bayesian cognitive model with arbitrary, debatable assumptions that came to our attention too late for inclusion here, by Gweon, Tenenbaum, and Schulz (2011), on infants' use of hypotheses about sampling techniques, is discussed at length in the supplement.)

At the extreme, when all other methods for explaining subject errors as optimal Bayesian reasoning have failed, theorists have in some cases decided that subjects were actually correctly answering a different question than the one the experimenter asked. For example Oaksford and Chater (Oaksford & Chater, 2009) proposed to explain errors in the well-known Wason card selection task by positing that the subjects were assuming the distribution of symbols on cards that would occur in a naturalistic setting; and arguing that, under that assumption, their answers were in fact optimal. At first glance, this seems to be offering a way of rescuing optimality, but in reality it just shifts the locus of non-optimality elsewhere, to the process of language comprehension.

Tenenbaum and Griffiths (Tenenbaum & Griffiths, 2001b) adopted much the same strategy in an analysis of the finding that subjects believe that the sequence of coin flips THTHTHHT is more likely than the sequence TTTTTTTT, asserting that what subjects are trying to say is, essentially, “given THTHTHHT the maximum likelihood hypothesis is that the coin is fair, whereas given TTTTTTT the maximum likelihood hypothesis is that the coin is biased.”

Although there may be instances in which subjects do genuinely misinterpret an experimenter’s questions, such explanations should be posited infrequently and must have strong independent motivation. Otherwise, resorting to such explanations risks further weakening the predictive value of the framework as a whole. A response that can be rationalized is not the same as a response that is rational.

Table 2: Features that have varied across probabilistic models of human cognition.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study** | **Domain** | **Probabilities** | **Derivation of those probabilities** | **Decision**  **Rule** |
| Hamrick et al. | Intuitive physics | Block position.  Form of distribution | Theoretically derived (Gaussian, corrected for interpenetration) | Maximum probability |
| Mean of position | Empirically derived |
| Standard deviation of position | Tuned post-hoc |
| Frank and  Goodman | Pragmatic reasoning with respect to communication | Probability that a particular word will be chosen for a given object | Information-theoretic model confirmed by experiment. | Proportional |
| Prior probability of that an object will be referred to, | Experimentally derived. |
| Griffiths and  Tenenbaum | Future predictions/”everyday cognition” | (Ii) Examples except waiting | Empirical distributions. | Median |
| (ii) Waiting example | inverse power law, tuned to fit subject responses. |
| Xu and Tenenbaum | Word learning | Priors on semantic categories and conditionals that an entity is in a category | derived from a complex model applied to experimentally derived dissimilarity judgments | Maximum probability |

**Caption**: Even in a relatively small sample of probabilistic models, a wide range of techniques are used in model construction, potentially chosen post hoc from a wider range of possible options. Many of the models would yield poorer fits if priors were derived differently, or different decisions rules were invoked (see, for example, discussion of Frank and Goodman in text).

Discussion

Advocates of the probabilistic approach have wavered about what it is that they are showing. At some moments they suggest that their Bayesian models are merely normative models, about what humans ought to do, rather than descriptive models about what humans are actually doing. Where the underlying mathematics is sound, there is no reason to question that modest interpretation. But there is also no reason to consider Bayesian models as a “null hypotheses” with respect to human psychology, in light of the apparent substantial empirical evidence that people sometimes deviate from normative expectation.

The real interest comes from the stronger notion that human beings might actually use the apparatus of probability theory to make their decisions, explicitly (if not consciously) representing prior probabilities, and updating their beliefs in an optimal, normatively sound fashion based on the mathematics of probability theory.

It would be too strong to say that humans never behave in apparently normative fashion; but it is equally too strong to say that they always do. As we have shown, people sometimes generalize in ways that are odds with correctly-characterized empirical data (poetry and film extrapolations), and in others generalize according to decision rules that are not themselves empirically sound (pragmatic reasoning task), or in ways at that are not empirically accurate (balance beam). The larger literature gives many examples of each of these possibilities, ranging from the underfitting of exponentials (Timmers & Wagenaar, 1977) to probability matching (West & Stanovich, 2003) to many of the cognitive errors reviewed by psychologists such as Kahneman and Tversky (Kahneman, 2003; Tversky & Kahneman, 1974, 1983).

We have also shown that the common assumption that "performance of a Bayesian model on a task defines rational behavior for that task" (Jacobs & Kruschke, 2011) is incorrect. As we have illustrated, there are often multiple Bayesian models with differing predictions based on differing assumptions; at most one of them can be optimal. Even though the underlying mathematics is sound, a poorly chosen probabilistic model or decision rule can yield suboptimal results. (In three of the examples we reviewed, subject performance that was actually *suboptimal* was incorrectly characterized as “optimal”, in part because of an apparent match between subject data and models selected post hoc that were Bayesian in character but incorrect in their assumptions.)

More broadly, probabilistic models have not yielded a robust account of cognition. Rather than a uniform architecture that is being applied across tasks, there is a family of different models, each depending on highly idiosyncratic assumptions tailored to individual tasks. Whether or not the models can be said to fit depends on the choice of task, how decision rules are chosen, and a range of other features. The Bayesian approach is by no means unique in being vulnerable to these criticisms, but at the same time, it cannot be considered to be a fully-developed theory until these issues are addressed.

The greatest risk here, we believe, is that probabilistic methods will be applied to all problems, regardless of applicability. Indeed, the approach is already well on its way to becoming a bed of Procrustes into which all problems are fit, even problems for which there are other much more suitable solutions.  In same cases, the architecture seems like a natural fit. The apparatus of probability theory fits naturally with tasks that involve a random process (Téglás et al., 2011; Xu & Garcia, 2008) and with many sensorimotor tasks (Körding & Wolpert, 2004; Trommershäuser, Landy, & Maloney, 2006), and AI systems that involve the combination of evidence. In other domains, however, such as intuitive physics and pragmatic reasoning, there is no particular reason to invoke a probabilistic model, and it often appears that the task has been made to fit the model. It is an important job for future research to sort between cases in which the Bayesian approach might genuinely provide the best account, in a robust way, and those in which fit depends on arbitrary assumptions. Ultimately the Bayesian approach should be seen as useful tool, not a one-size-fits-all solution to all problems in cognition. Griffiths et al’s (2012) effort to incorporate performance constraints such as a memory limitations could perhaps be seen as one step in this direction; another important step is to develop clear criteria for what would *not* count as Bayesian performance. (Another open question concerns development, with Xu & Kushnir (2013) suggesting that optimal, probabilistic models might be applied to children, while other studies such as Gutheil and Gelman, 1997 and Ramarajan, et al, 2012 suggesting some circumstances in which children too might deviate from optimal performance.)

The claims of human optimality, meanwhile, are simply untenable. Evolution does not invariably lead to solutions that are optimal (Jacob, 1977; Marcus, 2008), and optimality cannot be presumed in advance of empirical investigation. Any complete explanation must wrestle more seriously with the fact that putative rationality very much depends on what precise task subjects are engaging in, and offer a *predictive* account of which tasks are and are not likely to yield normative-like performance.

More broadly, if the probabilistic approach is to make a lasting contribution to our understanding of the mind, beyond merely flagging the obvious facts that people (1) are sensitive to probabilities, and (2) adjust their beliefs (sometimes) in light of evidence, its practitioners must face apparently conflicting data with considerably more rigor. They must also reach a consensus on how models will be chosen, and stick to that consensus consistently. At the same time, on pain of unfalsifiability, they must consider what would constitute evidence that a probabilistic approach was *not* appropriate for a particular task or domain; if an endless array of features can be varied in arbitrary ways, the framework loses all predictive value.

**Acknowledgements**

Thanks to Steven Pinker, Marjorie Rhodes, Gregory Murphy, Athena Vouloumanos, Michael Frank, Noah Goodman, and Fei Xu for many helpful criticisms and suggestions and to Michael Chagnon for assistance in data collection. Particular thanks to Tom Griffiths for providing the numerical data for the poetry estimation experiment.

**References**

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396.

Boseovski, J. J., & Lee, K. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology*, *42*(3), 500-13. doi:10.1037/0012-1649.42.3.500.

Bowers, J. S. & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated' subjects: Misconceptions about trajectories of objects. *Cognition*, *9*(2), 117-23.

Eberhardt, F. & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of bayesian models. *Minds and Machines*, *21*(3), 389-410.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc. Retrieved from Google Scholar.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. doi:10.1126/science.1218633

Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684. Retrieved from Google Scholar.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623-7. doi:10.1126/science.1223416

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767-73. doi:10.1111/j.1467-9280.2006.01780.x

Griffiths, T., Vul, E., & Sanborn, N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21(4), 263-268.

Gutheil, G., & Gelman, S. A. (1997). Children's use of sample size and diversity information within basic-level categories. *J Exp Child Psychol*, *64*(2), 159-74. doi:10.1006/jecp.1996.2344

Gweon, H., Tenenbaum, J.B., & Schulz, I.E. (2011). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, 107(20), 9066-9071.

Hamrick, J., Battaglia, P., & Tenenbaum, J. B. (2011). Internal physics models guide probabilistic judgments about object dynamics. *Proc. 33rd Conf. Cognitive Science Society.* Retrieved from Google Scholar.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161-6.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8-21. doi:10.1002/wcs.80

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *The Behavioral and Brain Sciences*, *34*(4), 169-88; disuccsion 188-231.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 1449-1475. Retrieved from Google Scholar.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237. Retrieved from Google Scholar.

Khemlani, S. S., Lotstein, M., & Johnson-Laird, P. (2012). The probabilities of unique events. *PLoS ONE*, *7*(10), e45975-. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pone.0045975

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244-247. Retrieved from Google Scholar.

Leary, M. R., & Forsyth, D. R. (1987). Attributions of responsibility for collective endeavors. Retrieved from Google Scholar.

Loftus, E. F. (1996). *Eyewitness testimony.* Harvard University Press. Retrieved from Google Scholar.

Marcus, G. F. (2008). *Kluge: The haphazard construction of the human mind.* Boston: Houghton Mifflin. Retrieved from Library of Congress.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 114-122. Retrieved from Google Scholar.

McNamara, J. M., Green, R. F., & Olsson, O. (2006). Bayes’ theorem and its applications in animal behaviour. *Oikos*, *112*(2), 243-251. Retrieved from Google Scholar.

Oaksford, M., & Chater, N. (2009). Précis of bayesian rationality: The probabilistic approach to human reasoning. *The Behavioral and Brain Sciences*, *32*(1), 69-84; discussion 85-120. doi:10.1017/S0140525X09000284

Ramarajan, D., Vohnoutka, R., Kalish, C., & Rhodes, M. (2012). Does preschoolers' evidence selection support word learning? *In Prep.*

*R*osenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638. Retrieved from Google Scholar.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173-220. Retrieved from Google Scholar.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481-520. Retrieved from Google Scholar.

Tenenbaum, J. B., & Griffiths, T. L. (2001a). Generalization, similarity, and bayesian inference. *Behav Brain Sci*, *24*(4), 629-40; discussion 652-791.

Tenenbaum, J. B., & Griffiths, T. L. (2001b). The rational basis of representativeness. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 1036-1041. Retrieved from Google Scholar.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-85. doi:10.1126/science.1192788

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054-9. doi:10.1126/science.1196404

Timmers, H., & Wagenaar, W. A. (1977). Inverse statistics and misperception of exponential growth. *Attention, Perception, & Psychophysics*, *21*(6), 558-562. Retrieved from Google Scholar.

Trommershäuser, J., Landy, M. S., & Maloney, L. T. (2006). Humans rapidly estimate expected gain in movement planning. *Psychological Science*, *17*(11), 981-988. Retrieved from Google Scholar.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-31. doi:10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293. Retrieved from Google Scholar.

West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Mem Cognit*, *31*(2), 243-51.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*(5), 440-445. Retrieved from Google Scholar.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 5012-5. doi:10.1073/pnas.0704450105

Xu, F. & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, *22*(1), 28-32.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245-72. doi:10.1037/0033-295X.114.2.245