Probabilistic models

David Danks

For: M. Colombo & M. Sprevak (Eds.), *Routledge handbook of the computational mind.*

1. A plethora of probabilistic models

The world is a fundamentally noisy and variable place: few events must occur; our measurements are rarely perfectly accurate; and relations are almost never deterministic in nature. Instead, there is uncertainty and error of various types in all our experiences, as shown by just the slightest reflection on everyday life. Sometimes, caffeine helps me to be more alert, but not always. Sometimes, my dog barks at strangers, but not always. Nonetheless, cognitive systems (including people) must be able to learn and reason appropriately despite this ineliminable noise and uncertainty. And in addition to variability in our experiences, human behavior is itself noisy and uncertain; people do not (and often should not) act identically in seemingly identical situations or contexts. Computational models of human cognition must have some way to handle all of the noise, uncertainty, and variability; many do so with probabilities, as the probability calculus is a standard computational framework for capturing and working with noise and uncertainty, whether in the world or the reasoner.[1] As one illustrative example, almost all theories of category judgments (such as "Is this a dog?") are probabilistic in nature: they allow for uncertainty in both the world—the same observation might sometimes be a dog, sometimes a wolf—and in the human categorizer—the same observation can probabilistically yield one of several possible judgments.

Although probabilities might be a standard tool for a computational cognitive model to capture noise and uncertainty, they nonetheless raise significant challenges for explanation and prediction.

[1] Probabilities are not the only way to address this issue—fuzzy sets (Zadeh, 1965) are another representational framework—but we focus on probabilistic models for reasons of space.

At a high level, the core underlying issue is that probabilistic models do not provide specific predictions for single cases or particular behaviors; instead, they only provide predictions about (features of) collections of behaviors. Almost any sequence of events is consistent with almost any probability distribution, although it might be highly improbable, and so our explanations and predictions do not operate in the usual ways. Instead, we need to rethink the explanations and predictions that these models provide. In this chapter, we consider these issues, as well as some of the novel benefits and advantages that probabilistic cognitive models can potentially provide beyond possible descriptive adequacy. Because the focus here will be on more conceptual issues, there will be few technical details. There are many useful introductions available elsewhere for readers interested in the precise mathematical formulations of probabilistic models in general (Koller & Friedman, 2009; Ross, 2009), and probabilistic cognitive models in particular (Chater, Tenenbaum, & Yuille, 2006; Perfors, Tenenbaum, Griffiths, & Xu, 2011). In addition, we will focus on cognitive models, rather than neurocomputational ones. Although many models of neural phenomena are probabilistic in nature (e.g., Doya, Ishii, Pouget, & Rao, 2007; Ganguli & Simoncelli, 2014), we will restrict our attention to more cognitive models (though many of the observations apply with minor adjustments to neurocomputational models).[2]

At a high level, probabilities can be incorporated into a computational cognitive model in two different, not mutually exclusive, ways. First, representations in the cognitive system can include or employ probabilities, where we take a very broad view of the notion of "representation." Any cognitive system must encode, whether explicitly or implicitly, key information about its environment, and these encodings will frequently involve probabilities to capture noise and uncertainty about the surrounding environments and contexts. For example, representations of

[2] We will also largely ignore debates about whether probabilities are subjective degrees of belief, physical propensities, limiting relative frequencies, or something else. In context, it is almost always clear how the probabilities are intended in these cognitive models.

causal structure are often modeled as probabilistic (causal) graphical models (Danks, 2014; Griffiths & Tenenbaum, 2005), which explicitly use a joint probability distribution to represent the noisy causal relations in the world. Or some Bayesian cognitive models represent theoretical knowledge as distinct hypotheses (perhaps probabilistic, perhaps deterministic) with probabilities that encode strength of belief (Griffiths, Kemp, & Tenenbaum, 2008; Perfors, *et al.*, 2011). A more implicit use of probabilities can be found in exemplar theories of categorization. These theories represent a category by a set of (definite, non-probabilistic) previously observed instances, and so appear to be non-probabilistic. However, those exemplars (plus a similarity metric) implicitly encode the probability of observing various types of individuals (Ashby & Maddox, 1993); that is, these categories actually correspond to probability distributions, even though they are not typically written in that way. In all of these cases, the cognitive model encodes or represents the world as a fundamentally noisy place; probabilities here are used to capture indeterminism in the environment, at least from our perspective.

Second, probabilities can be used in a computational cognitive model to capture noise and indeterminism in the cognitive agent herself. Experience, observations, and context rarely fully determine people's cognitive activity, at least at the level of our cognitive models. For example, our choices between two similar options will exhibit a degree of noise: given seemingly the same choice, we will sometimes pick option A and other times option B. Similar indeterminacy can be found in many other cognitive processes, and so our models of the agent's cognitive processes often include probabilities (even when the agent's representations are non-probabilistic in nature). In general, we can usefully distinguish between three types of cognitive processes, though no bright lines can necessarily be drawn to separate them: (i) learning; (ii) reasoning or inference; and (iii) acting or decision-making. A probabilistic learning process might yield different learned representations, even if identical observations and initial knowledge or prior beliefs are provided as inputs. A probabilistic

reasoning process might yield different judgments or beliefs, even given identical representations and context as input. A probabilistic decision-making process might yield different choices, even if given as input identical representations, beliefs, context, and goals. In each case, identical input to the process can yield different outcomes, and probabilities are used to capture this indeterminism.

Of course, probabilities can enter into a computational cognitive model at more than one place. For example, consider models of category acquisition and categorization—how we learn particular concepts, and then employ them for novel cases. In almost all cases, this cognitive process is noisy and indeterministic, and so should presumably involve probabilities. However, those probabilities can occur in representation (e.g., Rehder, 2003), learning (e.g., Love, Medin, & Gureckis, 2004), reasoning (e.g., Nosofsky & Palmeri, 1997), or more than one of the above (e.g., Tenenbaum & Griffiths, 2001). That is, we face an underdetermination problem: we know that probabilities have to appear *somewhere*, but we do not have the necessary data to determine whether they occur in representations, processes, or both. Often, the same behavioral phenomena can be modeled using (a) deterministic representations and probabilistic processes; (b) probabilistic representations and deterministic processes; or (c) both probabilistic representations and processes. Of course, most (complex) cognitive models face underdetermination challenges, but the problem here is even harder, as we do not even know what *type* of components (probabilistic vs. deterministic) should be employed in our model.

This introductory section has talked about computational cognitive models as though they apply to particular individuals; that is, cognitive models were discussed in the context of explaining the cognitive processes of particular individuals. In fact, though, many of our cognitive models are fundamentally ambiguous about whether they describe individuals or populations. In many contexts, this ambiguity is innocuous, but probabilistic models are not such a context. Suppose that we observe variability in behavior for a group of people who have all seemingly been exposed to the

same information (e.g., experimental stimuli in the lab). This variability could arise from everyone having the same probabilistic cognition, or from people having different deterministic cognition, or a mixture of the two. As a non-cognitive example, suppose that I flip many different coins and find approximately 50% heads, 50% tails. This "behavior" could arise at the population level because each individual coin is fair and balanced, or because half of the coins are two-headed and half two-tailed, or because we have a mix of these two extremal possibilities. More generally, any population-level probability distribution can be explained by probabilities *within* the individuals, or by probabilities *across* the individuals (or a combination). Perhaps we behave differently from one another because our cognition is fundamentally probabilistic, or perhaps because we have variability in our initial beliefs and subsequent experiences. The challenge for many of our computational cognitive models is that they describe average or population-level behaviors without explicitly stating whether the model is also an individual-level one. Often, it is implied that the models apply to particular individuals (not just the population), but that is frequently not explicitly stated. And in many cases, we lack the evidence to distinguish between the various possibilities, as we need repeated observations of each person in order to establish whether their particular cognition is probabilistic, and such repeated measures can be quite difficult and expensive to obtain. In the remainder of this chapter, we will see several places where this ambiguity—do the probabilities in the cognitive model capture within-individual or across-individual variability?—matters in the use, interpretation, and explanatory power of probabilistic cognitive models.

2. Explanation and prediction with probabilistic models

We begin by thinking about prediction using probabilistic models, as that is key to thinking about their explanatory power (as well as many other uses of probabilistic models). Importantly, probabilistic cognitive models will generally not predict any specific behavior at all, but rather a

range and likelihood of possible behaviors (regardless of where the probabilities are located in the model). These predictions can thus be quite difficult to assess or use, precisely because they are logically consistent with almost anything. Almost any sequence of data will be logically consistent with almost any probabilistic cognitive model, though the data might be quite unlikely. We thus need to rethink the exact content and target of our predictions and explanations.

It is perhaps easiest to see the issues by considering a non-cognitive example. Suppose that I am flipping a fair coin—that is, a coin that has a 0.5 probability of coming up heads. The predicted possibilities for flips of this coin include every possible sequence of heads and tails; some sequences might be exceptionally improbable, of course, but they are nonetheless possible. In the case of probabilistic cognitive models, almost any behavior will be predicted to be possible, though the model might predict that this behavior should be unlikely or rare. One reaction would be to conclude that probabilistic cognitive models are therefore untestable or useless, as they do not constrain the possibility space for behavior. This reaction is too quick, however, as we can instead shift to thinking about whether the observed behavior is likely or expected. Of course, we cannot test the likelihood of a single instance, and so we must also shift our focus from predicting a single behavior to predicting properties of sequences or collections of behaviors. This change raises anew the issue from the end of the previous section: if our focus is on collections of behavior, then we have to be very careful to distinguish between (a) collections formed from multiple behaviors by a single person; and (b) collections formed from behaviors by multiple people. Probabilistic cognitive models for (a) can be used to generate predictions for (b), but not vice versa. Hence, if our only observations are of type (b), then we will likely face additional underdetermination in terms of confirmation and plausibility.

With this understanding of the predictions of probabilistic cognitive models in hand, we can turn to the explanations provided by a well-tested, well-confirmed probabilistic cognitive model.

There are multiple explanatory virtues (as we will discuss in the next section), but we can first focus on the role of prediction in explanation. Predictive power is important because all theories of explanation hold that an explanans **S** should, in some sense, show why an explanandum *E* was expected, likely, inevitable, or otherwise followed naturally. Different theories of explanation provide different ways to explicate the idea of "following naturally," but all of those explications are connected in some way with predictive power. As a result, the approach that we employ for prediction in probabilistic cognitive models must also apply to the explanatory power of those models. In particular, probabilistic models cannot provide the same types of explanations, or same explanatory power, as deterministic cognitive models.

Consider some probabilistic cognitive model *M* and any arbitrary, though relevant, behavior *B*. As long as *M* assigns non-zero probability to *B*, then *M* can always give an "explanation" of *B*: the fact that *M* implies *B* is possible means that there is *some* sequence in *M* that results in *B*, and this sequence shows how *B* could have been produced (if *M* were true). But this means that the mere existence of an *M*-explanation is quite uninformative, since we know *a priori* that we will almost certainly be able to provide a story about how *B* could have been produced, regardless of what *B* turns out to be. And if a theory can "explain" any possible data, then it arguably provides no explanation at all. One natural response is to argue that *M* provides an explanation only if it shows that *B* is highly likely or highly probable. The problem, though, is that improbable things sometimes happen, and so this constraint implies that we will sometimes have no explanation for some *B* (i.e., the improbable ones). For example, a sequence of ten heads when flipping a fair coin is highly improbable—it will happen only around 0.1% of the time when one does ten coin flips—but if it does happen, then it would quite strange to say that we have no explanation at all.[3] More generally,

[3] In fact, *any* specific sequence of heads and tails will happen only 0.1% of the time when one flips a coin ten times, so we would actually have to say that we cannot provide an explanation for any

many probabilistic cognitive models predict that any particular, specific behavior *B* will be relatively improbable, even though they might well be able to provide a causal or mechanistic account of how *B* was generated.

At this point, there are two natural moves that one can make. First, we can change our understanding of the behavior to be explained. In the coin flipping case, any particular sequence is improbable, but sequences with certain shared features might be much more probable; for example, a sequence with five heads and five tails, regardless of order, occurs 24.6% of the time. Hence, we can perhaps save the requirement that an *M*-explanation should show how *B* is probable (or at least, not too improbable) by focusing on particular features of *B*, rather than *B* exactly. In the case of probabilistic cognitive models, this move typically requires shifting from explanations of a particular behavior *B* to explanations of *collections* of behaviors B_1, …, B_n. That is, *M* no longer provides an explanation of how a specific behavior, decision, or judgment resulted, but instead explains how features of a *distribution* of behaviors results, whether within a single person over time, or across a number of different people. These explanations of higher-level behavioral patterns are different than what we might have expected, but can be exactly what we want and need in certain contexts.[4] For example, if I am trying to understand causal reasoning, then I do not necessarily need to know how each particular causal judgment is generated, but only how they are usually generated, or the variability in how they can be generated, or the factors that are causally and/or explanatorily relevant to variation in those judgments. Individual people can be idiosyncratic in many different ways, and it might simply be unreasonable to think that we could give satisfactory explanations for how each specific behavior is generated; human cognition might simply be too complex a system. At the same

---

particular sequence, though as noted below, we could arguably explain certain properties of the sequence (e.g., proportion of heads being greater than, say, 0.4).

[4] As a non-cognitive example, note that this is exactly what we do in most thermodynamic models: we focus on predictions and explanations of properties of the distribution of particle locations, rather than specific particle locations.

time, we need to recognize that explanations of group-level phenomena or collections of behaviors (including those of the same person at different points in time) are importantly different from those that explain specific individual behaviors. We have changed our target, and so our explanations are arguably weaker in important ways. For example, they no longer explain any particular cognitive or behavioral event.

A second response is to shift away from asking whether *M* makes *B* probable or not, and instead focus on the sequence of events identified by *M* in its purported explanation of *B*. That is, we can require our *M*-explanations to provide an account of what actually happened to result in *B*. There are many debates about whether cognitive and neuroscientific explanations must be causal, mechanistic, or have some other shared feature (e.g., Craver, 2007; Kaplan & Craver, 2011; Lange, 2013; Ross, 2015). However, we do not need to engage with those debates here, as all of the parties agree that explanations identify a particular sequence of events that led to *B*. Those debates are about whether there are further constraints on that sequence, such as requiring it to be a causal sequence or mechanism. Regardless of that question, *B* presumably resulted from a sequence of events, and identification of that sequence provides one kind of explanation. Thus, a probabilistic *M* can perhaps provide a non-probabilistic explanation of *B*. Unfortunately, as noted earlier, we know *a priori* that we will almost always be able to postulate *some* sequence of events in *M* that would lead to *B*. The key question for explanations of this type is whether the postulated sequence actually occurred, and that determination requires that we observe much more than just *B*. This second strategy—shift to focusing on actual sequences of events—might be the right one in some cases, but comes at a cost: we only have grounds to believe those *M*-explanations about how *B* actually resulted if we have much more information about the particular case. The mere observation of *B* is clearly not sufficient, since we can almost always generate a "how possibly" *M*-explanation.

The overall message is that probabilistic cognitive models generally provide explanations of how some behavior resulted only if we (i) weaken our expectations by shifting to features of collections of behavior (by the same or different individuals); or (ii) strengthen our measurement capabilities by observing intermediate states or events that culminated in the behavior. Given this choice, we might instead pursue a completely different response by changing the desired type of explanation to account for *why* the behavior occurred (rather than *how*). In particular, many probabilistic cognitive models have been offered as "rational" or "optimal" models that can tell us why some behavior occurred, even if nothing can be said about how it was generated. For example, a *rational* model of categorization (Anderson, 1991; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Tenenbaum & Griffiths, 2001) aspires to explain people's category judgments by showing it to be optimally correct behavior. These theories do not specify the processes or structures by which these judgments were produced, but they can nonetheless explain why people act as they do: namely, people are trying to succeed at the task of categorization, and these responses are the right way to do that. The explanation here is analogous to what one might say when asked to explain why a calculator reads '17' when '8+9=' is entered: namely, that's the right answer, and properly functioning calculators are designed to give the right answer. This explanation gives us no information about how the calculator functions, but it nonetheless can explain the calculator's "behavior." Of course, human cognition is not necessarily "designed" like a calculator is, and so we must provide additional information (Danks, 2008). Nonetheless, this different type of explanation—a why-explanation rather than how-explanation—is another response to the difficulty of explaining human behavior.

Why-explanations are not restricted to probabilistic cognitive models, but they are particularly common for those models, partly for reasons that we explore in the next section. For now, we focus on the requirements and explanatory power of these why-explanations. To have a full why-explanation, we need to show not only that the behavior is optimal for human cognizers in these

environments, but also that people act in this way *because* the behavior is optimal (Danks, 2008). The second requirement is crucial, as optimality alone does not tell us why the behavior occurred if, in fact, that optimality played no role in leading to the behavior. In the calculator case, the why-explanation depends partly on the calculator being correctly designed; a similar claim is required for why-explanations of human behavior. Of course, we do not need to have a full causal story about the role of past (optimal) performance. For example, it is sufficient to show that there are ontogenetic or phylogenetic pressures that will push people to act more optimally. And given a demonstration of optimality and its cognitive relevance, then not only do we arguably know why some behavior occurred, but we can also predict what would happen if the environment or task shifted (assuming the individual had time to learn and adapt).

This why-explanation is limited in certain important ways. For example, and in contrast with a causal-mechanical how-explanation, we can make only limited predictions about what might happen if the cognitive system is damaged or altered in some way. We can presumably expect that it will be different in *some* way, but we cannot predict how exactly it will change, nor whether it will be able to recover or adapt to this damage. We also do not avoid the problem of predicting single cases: if the optimal behavior is to act probabilistically (as in, for example, certain foraging situations), then we will still have to shift our explanandum to properties of collections of behaviors. Nonetheless, these why-explanations do represent a qualitatively different type of explanation from the usual causal-mechanical-computational ones found in cognitive science.

3. Explanation beyond prediction

One key feature of explanations is that they show why or how something occurred, but there are plausibly other explanatory virtues or functions. In particular, explanations are often thought to play a unifying role (Kitcher, 1981), though the nature of this unification is not always clear. In the case

of probabilistic models, the unification function is often touted as an important aspect that speaks in favor of the models. These arguments all begin with the observation with which this chapter started: the world is a fundamentally noisy and uncertain place (from our perspective). The proponents of probabilistic models then typically argue that the probability calculus, or Bayesian updating, or some other probabilistic model is the normatively correct way to handle noise and uncertainty (e.g., Chater, *et al.*, 2006; Chater & Oaksford, 2008; Oaksford & Chater, 2007). Thus, these arguments claim that probabilistic cognitive models provide explanatory unification in virtue of being the (purported) correct way to handle a world like ours. The shared language of probabilities in all of these models of diverse cognitive phenomena provides a further unification: they are all instances of probabilistic inference, reasoning, specification, etc., and so these cognitive processes and behaviors are just different manifestations of the same type of theoretical "machinery" (leaving aside the question of whether they share any neural "machinery").

There is a sense in which the conclusion of these arguments is correct, as our cognition surely must be robust in the face of various types of noise, uncertainty, or indeterminism. It would be bizarre if our cognitive processes had no way of representing and responding (perhaps implicitly) to this variability. And to the extent that we think that different cognitive phenomena do involve similar types of processes or representations, we should favor model-types that are widely successful. So to the extent that probabilistic cognitive models have significant, widespread *descriptive* explanatory success, then we might hope that we can develop an argument that future probabilistic models should be judged as more plausible. However, this unificationist argument makes no reference to rationality or optimality claims, and so we must provide further argument that rationality or optimality considerations provide a further (explanatory) reason to favor probabilistic cognitive models, rather than ones based in other non-deterministic processes.

The standard arguments for why-explanations in probabilistic cognitive models depend on the claim that all rational, optimal, or normative models *must* be probabilistic; in particular, they must satisfy the probability calculus. This claim justifies assertions that probabilistic models are the "correct" or "appropriate" way to handle uncertainty, which thereby privilege those models (when they are approximately descriptively correct). There are many different defenses of this claim in the literature (many collected in Oaksford & Chater, 2007). For example, Dutch book arguments show that failing to conform to the probability calculus can lead to decisions that are individually sensible (from the decision-maker's point of view) but are jointly guaranteed to end badly. Or convergence arguments show that no method of changing one's beliefs can consistently do better than if one changes strengths of belief according to the probability calculus. Many of these arguments are deployed specifically in favor of Bayesian models—that is, models in which belief change or inference occurs through conditionalization as given by Bayes's Rule—but they often are appropriate for probabilistic models more generally. The details also obviously can matter in these arguments for the crucial claim, but the key here is simply that they all aim to establish strong, perhaps even identity, relations between the class of probabilistic models and the class of rational/optimal models.

However, there is an issue with the way that these arguments are used. In every case, the arguments show (at best) that probabilistic models, reasoning, or updating are *one* good way to handle uncertainty, not that they are the *only* or *uniquely* good (or rational, or optimal, or correct) way (Eberhardt & Danks, 2011). More specifically, "probabilistic" and "rational" are theoretically independent notions: one can have probabilistic, non-rational models, and also non-probabilistic, rational models. Although probabilistic models are often rational or optimal, they are not privileged in that way. As just one example, consider the cognitive task of learning from experience. There have been numerous arguments that Bayesian conditionalization is the rational way to learn—that is,

given a new piece of evidence, the changes in one's probabilities over various options (i.e., learning from a probabilistic perspective) should change in accordance with Bayes's Rule (Teller, 1973, 1976). The normative force of these arguments arises from Bayesian conditionalization ensuring probabilistic coherence over time, or consistency of plans in light of new information, or convergence to the truth (when it is learnable), or other such desirable features. But in every case, there are alternative methods—sometimes, infinitely many such methods—that also satisfy that desideratum (Douven, 1999; Eberhardt & Danks, 2011). Bayesian conditionalization is normatively defensible, but not uniquely normatively privileged, compared to other methods for shifting belief, or other (often more qualitative) representations of uncertainty.

This story repeats itself for essentially every argument in favor of the rationality of probabilistic models and methods: they are normatively defensible, but not normatively unique. Moreover, explanatory unification (of the sort proposed at the beginning of this section) depends on uniqueness, not simply defensibility. Probabilistic models and methods were argued to provide some extra explanatory power that goes beyond "mere" descriptive adequacy, but the additional explanatory power depends on the number of alternatives. If probabilistic models and methods are only one of many possibilities, then we have only a very weak normative explanation of why the brain/mind employs them (if it does). We cannot claim that these models are inevitable because "a rational agent couldn't have done otherwise," precisely because there are many different things that a rational agent could do instead. That is, the question "Why probabilistic models?" cannot be answered with "Because they are inevitable for rational agents," despite suggestions to the contrary from proponents of those models.

Despite these issues, there is still an important sense in which probabilistic models and methods can provide a type of explanatory unification, though one grounded in their descriptive rather than normative virtues. These theories employ a common template or schema for the specification of the

model, methods, and techniques (Danks, 2014, ch. 8). We can thus understand the mind as consisting of many distinct instantiations of the same underlying type of representation or process, such as joint probability distributions or Bayesian updating processes (Colombo & Hartmann, 2017). To the extent that we expect there to be similarities within the mind, the shared schema of a probabilistic method or model implies that the collection of probabilistic models has greater explanatory power than the "sum" of the individual model's explanatory powers. That is, the shared probabilistic schema implies mutually reinforcing support,[5] at least to the extent that we expect that different aspects of the mind/brain should have some degree of similarity. One might worry about this last qualifier, as there are many arguments that the mind/brain should and does exhibit substantial modularity, and we might have no particular reason to think that modules share a model- or method-schema (Carruthers, 2006; Fodor, 1983; Tooby & Cosmides, 1992). However, we also have no reason to think that modules *cannot* have a shared schema, as apparently module-specific phenomena can instead arise because of distinct prior knowledge, experience, or expectations (Samuels, 1998). General arguments for modularity do not speak directly against "schema-based" explanatory unification. We thus find that probabilistic models do arguably have some (potential) additional explanatory power if they are as widespread as proponents claim, but it is based in their descriptive similarity, not a shared normative base.

4. Conclusion

We live in a noisy, uncertain world, and probabilistic models, methods, and reasoning are a natural way to tackle such environments. We should thus be unsurprised that probabilistic models are ubiquitous in modern cognitive science: they are found in models of essentially every area of the mind/brain, from very early perception (Ganguli & Simoncelli, 2014; Lee & Mumford, 2003), to

[5] This interdependence can be made precise in terms of intertheoretic constraints (Danks, 2014).

both simple low-level (Courville, Daw, & Touretzky, 2006; Xu & Tenenbaum, 2007) and complex high-level (Chater, *et al.*, 2006; Oaksford & Chater, 2007) cognition, to motor activity (Kording & Wolpert, 2006; Wolpert & Kawato, 1998). They have been employed to understand even phenomena that are sometimes thought to be non-computational, such as emotions (Seth, 2013). At the same time, there are very real challenges in understanding the explanations that such models and methods provide. Almost any behavior is consistent with almost any (plausible) probabilistic cognitive model, and so many of the standard theories of prediction and explanation do not apply. Instead, we must shift how we think about explanation with these models. Instead of explaining a single particular instance, we can: explain features of the collections of phenomena (in individuals or groups); or collect additional measures that ground the particular explanation; or shift to providing why-explanations rather than how-explanations. Each of these strategies has been employed with probabilistic cognitive models, thereby enabling widespread use of these powerful types of models.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Carruthers, P. (2006). *The architecture of the mind.* Oxford: Oxford University Press.

Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind: Prospects for Bayesian cognitive science.* Oxford: Oxford University Press.

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

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *British Journal for the Philosophy of Science, 68*(2), 451-484.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience.* Oxford: Oxford University Press.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 59–75). Oxford: Oxford University Press.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models.* Cambridge, MA: The MIT Press.

Douven, I. (1999). Inference to the best explanation made coherent. *Philosophy of Science*, *66*, S424–S435.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.). (2007). *Bayesian brain: Probabilistic approaches to*

neural coding. Cambridge, MA: The MIT Press.

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

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: The MIT Press.

Ganguli, D., & Simoncelli, E. P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation*, *26*(10), 2103–2134.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational cognitive modeling*. Cambridge: Cambridge University Press.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*, 507–531.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: The MIT Press.

Kording, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science*, *64*, 485–511.

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

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 1141–1159.

Ross, L. N. (2015). Dynamical models and explanation in neuroscience. *Philosophy of Science*, *82*(1), 32–54.

Ross, S. M. (2009). *Introduction to probability models* (10th ed.). Academic Press.

Samuels, R. (1998). Evolutionary psychology and the massive modularity hypothesis. *The British Journal for the Philosophy of Science*, *49*(4), 575–602.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573.

Teller, P. (1973). Conditionalization and observation. *Synthese*, *26*, 218–238.

Teller, P. (1976). Conditionalization, observation, and change of preference. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 1, pp. 205–259). Dordrecht: Reidel.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L.

Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). Oxford: Oxford University Press.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*, 1317–1329.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Zadeh, L. (1965). Fuzzy sets. *Information and Control, 8*, 338-359.