Don't Trust Fodor's Guide in Monte Carlo:

Learning Concepts by Hypothesis Testing Without Circularity

to appear in *Mind & Language*

Michael Deigan
Rutgers University

> HOW TO GET ABOUT. The American Express and Cooks have circular bus tours of the region.
>
> [Eugene] *Fodor's Guide to Europe, 1967*

Can concepts be learned? Jerry Fodor argued that they cannot be, that the very idea of concept learning is "per se confused".[1] Concepts may be innate, they may be acquired through various non-rational processes that we shouldn't count as learning, but, according to Fodor, they cannot be learned.

Here's his argument:

(1) If there were concept learning, it would have to be by a process of hypothesis testing.

(2) Concepts cannot be learned by hypothesis testing, on pain of circularity.

Therefore,

(3) There can be no concept learning.

In broad brushstrokes, there are two options for resisting this argument: deny premise (1) or deny premise (2).

[*]For helpful discussions, thanks to Christopher Blake-Turner, Kirstine la Cour, Keith DeRose, Daniel Ferguson, Daniel Greco, Joshua Knobe, Joanna Lawson, Jason Stanley, Zoltán Gendler Szabó, Nadine Theiler, and a reviewer for *Mind & Language*.

[1]See Fodor (2008, Ch. 5), as well as Fodor (1998, Ch. 6). In Fodor (1975), Fodor had made a similar argument towards the conclusion that most of our concepts are innate. But in the later work I will be discussing, he tempered this conclusion, allowing that many concepts may be acquired in some way, just not through learning. The earlier argument also turned on the claim that most of our concepts are atomic, but the later arguments make no use of this.

Denying premise (1) is a popular strategy. According to many, we can learn concepts in ways that are not any kind of hypothesis testing. Some would grant Fodor that these are not really rational processes—whatever exactly that amounts to—but reject his assumption that learning must be a rational process.[2] Others hold that some such learning processes *are* rational processes, and so reject (1) even granting this assumption about learning.[3] I find this latter variant of the strategy congenial, and defend it in other, in-progress work.

What I want to do here, though, is grant premise (1) and take the less popular route of denying premise (2). A few others have attempted this, but I will be showing a new way it can be done.[4] It will involve putting two ideas together. The first is an idea from the metaphysics of ability that goes back at least to Kenny (1975): for an important sense of 'ability', actuality does not entail ability. I'll use this to argue that there is a gap in Fodor's defense of premise (2). The second idea allows us to exploit this gap, showing how concepts can be learned by hypothesis testing in a non-circular way. The idea, which comes from recent computational cognitive science, is that the kind of hypothesis testing involved in human learning is stochastic, involving generative random sampling, as in Monte Carlo methods used to approximate Bayesian inference. I'll call this approach to denying premise (2) the *Monte Carlo Way*.

1. Fodor's Circle

To start, we need to see how Fodor defends premise (2). Why does he think learning concepts by hypothesis testing would be viciously circular?

[2]For responses that explicitly deny that learning must be a rational process, see Margolis and Laurence (2011, p. 518) and Sundström (2019).

[3]See Sterelny (1989), Weiskopf (2008), Margolis and Laurence (2011, p. 519), Carey (2009), Carey (2011, p. 162), and Carey (2015).

[4]See Margolis and Laurence (2011) and Buijsman (2019) for other arguments against premise (2).

Hypothesis testing is just standard inductive inference. One begins with some candidate hypotheses (coming from who knows where—their origin is beyond the scope of inductive logic, or indeed any kind of logic, as far as Fodor is concerned). Then one tests these hypotheses on the basis of how well they conform with one's experiences, perhaps by eliminating them until one is left with just a single hypothesis, perhaps by assigning prior conditional probabilities and updating by conditionalization, perhaps by some other algorithm. Fodor does not go into the details about how this process might actually work, presumably because he thinks his argument will go through no matter what the details are.

Concepts cannot be learned by any kind of hypothesis testing, Fodor claims, because this would involve a vicious circle. Here's the argument.[5]

(i) To learn a concept C by hypothesis testing, you must consider a hypothesis about c-hood, which can then be confirmed by experience.

But (ii) if you consider a hypothesis about c-hood, you must already be able to think thoughts about c-hood.

So (iii) you must already possess C in order to do the relevant hypothesis testing.

So (iv) learning a concept by hypothesis testing presupposes that you already have the concept.

But (v) *learning* a concept would require that you did *not* already have the concept prior to the learning.

So (vi) hypothesis testing cannot be a way of learning a concept.

This is *Fodor's Circle*.

We might try to get out of Fodor's Circle by denying (v), along the same lines as Plato's response to Meno's Paradox, as some interpreters read it.[6]

[5] See Fodor (1998, pp. 123–124) and Fodor (2008, pp. 138–140).
[6] See, e.g., Fine (2014, Ch. 4).

On this view, one might really learn C through hypothesis testing even though one had the concept all along. But this would amount to challenging merely the letter of Fodor's conclusion rather than its substance.

A more promising way out is to reject (i). To do this, one would need to describe a way of testing hypotheses without having the hypothesis in mind, as Margolis and Laurence (2011) do, or a means of learning a concept C by hypothesis testing without confirming hypotheses about c-hood, as Buijsman (2019) does. No doubt such proposals will be controversial, but I myself do not wish to stir up any controversy about them. If rejecting (i) works, so much the worse for Fodor's argument.

The way out that I want to explore, though, does not need to deny either (i) or (v). Instead, it denies (ii).

2. Ability, Actuality, and a Gap in the Circle

Fodor's Circle turns on two assumptions about abilities. One is that there is a connection between abilities and concept possession, which Fodor makes explicit:

> **Ability Implies Concept Possession (AICP)**
> "A sufficient condition for having the concept C is: being able to think about something *as (a) C* (being able to bring the property C before the mind as such, as [Fodor] sometimes put[s] it)."
>
> Fodor (2008, p. 138, Fodor's emphasis).

AICP is required to make Fodor's Circle work, since it is what allows us to infer (iii) from (ii). If one could have an ability to think of something as a c without having the concept C, then one could test the relevant hypothesis without already having the concept, so the threat of circularity would dissipate. But Fodor takes AICP to be self-evident, and I am happy to grant it.

The second assumption, which is left implicit, is about how ability relates to actuality.

> **Actuality Implies Ability (AIA)**
> If someone actually ϕ's, then they had the ability to ϕ when they ϕ'd.

Fodor needs to assume AIA, since otherwise there would be no reason to accept (ii), which is just an application of AIA to abilities to think.

Fodor's Circle involves inferring from the fact that someone actually considers (and so thinks) a hypothesis involving the concept C to the conclusion that that person already had the concept C. For this inference to work, we need to be able to infer from someone's ability to think a thought to their possession of the relevant concept, as AICP allows. But we also need to be able to infer from the fact that someone *actually* thought a hypothesis to the conclusion that they *had the ability* to think it. This is what AIA allows when applied to abilities to think. Without it, or something very much like it, Fodor's Circle will have a gap, since the learner might actually consider a C-involving hypothesis without having the ability to think C-involving thoughts, in which case AICP would not kick in to get us to the conclusion that they already possess C.

Perhaps Fodor didn't highlight AIA because he didn't notice he was relying on it, or perhaps he took it to be so self-evident that it didn't need to be mentioned. Some others have taken AIA to be a clear conceptual/analytic truth: "of course", says J. L. Austin, "it follows merely from the premiss that he does it, that he has the ability to do it, according to ordinary English" (Austin 1956, p. 175). Suppose we're debating about whether Annika can sink this putt, for example. If she misses, this might not settle the question—perhaps she has the ability, but something went wrong in this case. However, if she in fact sinks it, this would seem to conclusively show that she did have the ability to sink that putt. So it may seem that AIA is on firm ground.

Moreover, it is natural to think that statements of abilities are equivalent to corresponding modal statements using *can*.

(1) a. Annika is able to sink this putt.
b. Annika can sink this putt.

And it seems plausible to understand the *can*'s of such statements as possibility modals with some kind of metaphysical flavor.[7] But all it takes to make a possibility modal of this kind true is the existence of a single accessible world where the prejacent is true. And since the actual world should always be an accessible one when we're dealing with metaphysical modality (actuality should imply possibility), truth of the prejacent in the actual world will suffice for truth of the modal sentence, and hence for the equivalent ability claim. In other words, a natural account of the meaning of ability statements lends further support to AIA.

However, things are not so rosy for AIA, as metaphysicians have long observed, going back at least to Kenny (1975). At least on one ordinary understanding of *is able to* (and the relevant sense of *can*), actuality does *not* imply ability. To see why, consider the following cases.

> **Lucky Hole-in-one**
> Bob is a terrible golfer, often badly messing up even easy shots. On this occasion, though, he luckily swings with decent form, then a lucky bounce off a tree and unexpected gust of wind carries his ball into the hole for a hole-in-one. Though Bob in fact hit a hole-in-one, he did not have the ability to hit a hole-in-one.

> **Random 58-ball**
> As part of an assignment for her statistics class, Cindy is drawing balls from a bag. There are 10,000 balls in the bag, each labelled with a different numeral from 1 to 10,000. Cindy does not have a special

[7] See Hilpinen (1969), Kratzer (1977), and Kratzer (1981).

> technique which allows her to draw the ball labeled '58', and so does not have the ability to draw the 58-ball. She shakes up the bag, draws a ball at random and, it turns out, draws the 58-ball. However, this does not mean she had the ability to draw the 58-ball after all.

Kenny (1975, p. 136) gives the example of a hopeless darts player who hits the bullseye once in his life, despite lacking the ability to hit the bullseye, as well as his own inability to correctly spell 'seize', which he actually spells correctly about half the time. There will be potential counterexamples to AIA, Kenny observes, "whenever it is possible to do something by luck rather than skill" (*ibid*.).

Cases like these—along with a slew of other arguments—have been used to show that a straightforward possibility modal analysis of ability modals cannot be right. At least some abilities require something besides a single possibility where the relevant event happens, so we need to analyze abilities (and statements about abilities) in some other way.[8] So what does it take to have an ability? This remains a contentious issue. Analyses have been proposed in terms of conditionals, other kinds of modality, generics, dispositions, and combinations thereof. We need not decide between these for present purposes. What's important is that such cases show that at least for one ordinary sense of 'ability', one might actually do something without having the ability to do that thing. For this kind of ability, where ϕ-ing depends too much on luck or randomness, AIA fails.[9]

Now we can see that there is a gap in Fodor's Circle, and so also in his argument against the possibility of concept learning: Fodor implicitly relies on AIA, but AIA is false. We can also see, at least schematically,

[8]Besides Kenny (1975), see Mele (2003), Vihvelin (2013, Ch. 1), Maier (2015), and Mandelkern, Schultheis, and Boylan (2017), though see also Schwarz (2020) for a recent defense of a version of the possibility modal analysis. For an argument that possibility is not necessary for ability, which would be another blow to the possibility modal analysis, see Spencer (2017).

[9]There may be other senses of 'ability' on which AIA holds, an issue which we will discuss in §4.4.

how concept learning by hypothesis testing could be possible. One might actually consider some hypotheses involving C while lacking the ability to think C thoughts. In which case one may not already possess C, even while testing a hypotheses that involve it. And through confirmation of such hypotheses, one may come to possess C.

The mere schematic possibility of concept learning is interesting. It shows us that Fodor has not in fact established that "it's true and a priori that the whole notion of concept learning is per se confused" (Fodor 2008, p. 130). And this is so even if we grant him his premise (1)—that concept learning would have to be done by hypothesis testing—and his premise (i)—that to learn C by hypothesis testing, one would need to consider a hypothesis about c-hood. In thinking about how to respond to Fodor's argument, it is significant that we can grant even what has been most contested but still resist his conclusion.

But beyond this negative point, it is not yet clear how seriously we should take the possibility of concept learning by the route just described. After all, the exceptions to AIA seem to fall within a fairly limited class. Usually doing something does suffice for showing one has the ability. It seems only to be where there is significant luck or randomness involved that AIA fails. Is there reason to think that there is a possible method of induction that could involve one actually considering hypotheses without having the abilities to think those hypotheses? And even if there is some such possible method, is there any reason to think that humans ever learn concepts in something like this way?

In the next section I'll argue that a mainstream line of work in cognitive science shows that the answer to both of these questions is "Yes". We should take the possibility of concept learning by AIA-violating hypothesis testing seriously, not just as a way to poke a hole in a famous argument of Fodor's, but as a way to make sense of how we may in fact acquire concepts through a rational process.

3. The Monte Carlo Way Through the Gap

3.1 Bayesians and their Generative Sampling Algorithms

On Bayesian approaches in cognitive science, various cognitive processes are cast as problems of inductive inference, where the aim of the processes is to approximate a Bayesian ideal.[10] According to this ideal, one starts with some prior subjective probability distribution $P(h)$ over hypotheses in a given hypothesis space $\mathcal{H}$, as well as likelihoods $P(d|h)$ specifying how likely the observation of some data d is supposing h is true. Then, as one makes observations, one updates these prior probabilities to get posterior probabilities via conditionalization, in accordance with Bayes' theorem, which require that the posterior probability of h is proportional to the product of its prior probability and its likelihood, relative to the sum of the products and likelihoods for all the hypotheses $h' \in \mathcal{H}$:

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

From perception to language learning, motor control to emotion recognition, there are now Bayesian models that have reached a high level of rigor and empirical support.[11] Of particular relevance to us are the impressive advances in understanding concept learning using Bayesian models,

[10] For overviews, see Griffiths, Kemp, and Tenenbaum (2008) and Perfors, Tenenbaum, Griffiths, et al. (2011). For more philosophically oriented discussions, see Rescorla (2015), Rescorla (2016), and Icard (2018).

Two clarificatory observations: first, cognitive scientists can use Bayesian methods for data analysis and modelling without committing to Bayesian models in the relevant sense, just as population biologists can use Bayesian methods without taking populations to be doing Bayesian inference. Second, though there are important commonalities, most of the Bayesians I will be discussing are not advocates of the Predictive Processing models defended by Friston (2010), Hohwy (2013), and Clark (2016).

[11] See, e.g., Perfors, Tenenbaum, and Regier (2011), Xu and Tenenbaum (2007), Kersten, Mamassian, and Yuille (2004), Körding and Wolpert (2004), Ong, Zaki, and Goodman (2015).

primarily due to Joshua Tenenbaum and his collaborators, to which we will soon return.[12]

Bayesian approaches have long been among the main contenders for thinking about norms of inductive inference among statisticians, philosophers of science, and epistemologists.[13] The Bayesian turn in cognitive science is more recent, largely because performing the relevant computations used to be intractable. This meant Bayesian models were neither useful to scientists for making quantitative predictions nor attractive as hypotheses for how cognitive processes actually work.

This changed with the advent of computationally tractable algorithms for approximating Bayesian inference, together with an increase in computational resources. Cognitive scientists now find themselves with tools that allow them to easily determine quantitative predictions of Bayesian models of a variety of cognitive processes, and have often found these predictions to be borne out.[14] And this has not been merely an advance in modellers' tools. Interestingly, certain quirks of these algorithms—aspects in which they systematically diverge from Bayesian ideals—have been found to be reflected in human behavior. Many cognitive scientists have accepted a natural explanation of this: that human minds use some such methods in perception, judgement, inference, learning, and so on. More specifically, many Bayesian cognitive scientists hold that human brains compute probabilities using *sampling* methods to approximate ideal Bayesian inference.[15]

[12]This work goes back his dissertation, Tenenbaum (1999). Milestone publications include Tenenbaum et al. (2011), Lake, Salakhutdinov, and Tenenbaum (2015), Lake, Ullman, et al. (2017).

[13]For overviews, see Earman (1992), Easwaran (2011a), Easwaran (2011b), and Sprenger and Hartmann (2019).

[14]On the increase of Bayesian modelling in cognitive science due to computational tractability, see Lee and Wagenmakers (2013, p. 7).

[15]See Tenenbaum (1999), Griffiths, Vul, and Sanborn (2012), Denison et al. (2013), Vul et al. (2014), Sanborn and Chater (2016), Icard (2016), Thaker, Tenenbaum, and Gershman (2017), and Lloyd et al. (2019). Proponents of the Predictive Processing variant of Bayesianism

How do these sampling methods work? There are many varieties, full specifications of which would require getting more involved than is necessary for our purposes. Instead, I'll give an informal overview of one popular class of sampling methods: Markov chain Monte Carlo methods.[16]

In general, *Monte Carlo* methods are algorithms that use random sampling to approximate exact solutions to various problems. Suppose, for example, we want to find an integral of some complicated function of several variables. Rather than computing it analytically, which may be impossible, or through some deterministic approximation algorithm, which may be intractable, we can approximate the solution by drawing many random samples from a uniform distribution over the domain, evaluating the function at those samples, taking the average of these results and multiplying it by hypervolume of the domain. This will approximate the solution, and the more samples we take, the more accurate the approximation is likely to be. The result is a method for determining the integral to an arbitrary degree of precision in a relatively tractable way.

For some problems, as in the integral case, the random sampling can be done from a uniform distribution, or some other standard distribution that can be straightforwardly sampled from. In other cases, however, this won't do. In particular, this won't work for estimations of Bayesian posterior probabilities. Intuitively, to estimate a posterior probability or likelihood, we would want to sample from the posterior distribution itself—so that we are more likely to draw samples from regions where the posterior probability is higher. But we usually don't have any way of doing this, assuming that the posterior probability is difficult to compute directly.

Fortunately, there are Monte Carlo methods that circumvent the need

claim the human mind uses a different kind of approximation method—variational methods—to approximate Bayesian inference. See Sanborn (2017) for a comparison.

[16]Many of the articles cited above provide relatively accessible overviews of these and other sampling methods. For more detail, see Kruschke (2010, Ch. 7), MacKay (2003, Ch. 29), and Andrieu et al. (2003).

for directly sampling from the target distribution. *Markov chain Monte Carlo* (MCMC) methods, for example, do this by by taking a biased random walk through the sample space. So long as we can compare the relative posterior likelihoods of our samples, we can keep the higher likelihood samples, throw out a portion of the lower likelihood samples, and 'explore' the space in a way that spends more time in higher likelihood regions, resulting in a collection of samples whose frequencies approximate the posterior probabilities. Among other applications, MCMC methods can allow for approximations of various features of an ideal Bayesian posterior, even when computing these features analytically is intractable. For this reason, cognitive scientists have proposed that MCMC methods are used in a variety of cognitive processes that involve hypothesis testing. In §3.3, we'll discuss a concrete example of such a proposal for concept acquisition.

For the issue at hand, the crucial fact about these algorithms for hypothesis testing is that they do not proceed by considering and evaluating all of the hypotheses in the hypothesis space, but instead proceed by *generative random sampling*. Since the hypothesis spaces in question may be infinite (or finite but enormous), going through all the hypotheses cannot be done, so these methods generate and evaluate only a relatively small collection of samples from the hypothesis space in a way that manages to be representative of the whole space. And not only do they generate merely a very sparse portion of the space, they generate the samples randomly. And this kind of randomness is just what we need for an AIA-violating process. Generative random sampling approaches to inductive inference, we'll see, show that the way of non-circular concept learning by hypothesis testing suggested in the previous section is not a mere schematic possibility useful for making a philosophical point, but a real candidate for how at least some concept learning actually works.

Before spelling out how concept learning could work by generative random sampling, though, two remarks are in order. First, we might want

to ask whether using such algorithms *really* counts as hypothesis testing. I assume this particular kind of hypothesis testing is not one that Fodor had in mind. Nevertheless, this is how hypothesis testing is often done in the quantitative sciences, and how many psychologists think it is done in much of human cognition. We could read Fodor's premise (2) in such a narrow way that it excludes MCMC methods, and may thereby avoid the impending counterexamples. This is an uninteresting response, however, since it just moves the argumentative bump in the rug, making the present proposal an objection to premise (1) rather than premise (2), in a way that obscures important differences between the Monte Carlo Way and other ways of objecting to Fodor's argument.

Second, it is worth noting that there have been various objections from both philosophers and psychologists to Bayesian models in cognitive science, at least when such models are construed in realistic ways.[17] There have also been replies to these criticisms.[18] Strictly, these debates do not affect my primary aim, which is to refute Fodor's claim that concept learning is impossible. Since the criticisms of Bayesian modelling do not purport to show that it is impossible that the mind works in these ways, they do not undermine my use of these theories to show that concept learning by hypothesis testing is possible. Nevertheless, my argument will be more interesting if the Bayesian program is on the right track and an approximately Bayesian way of learning concepts is how we learn at least some of our concepts.[19] In that case, we will have an idea of not just how

[17]See Jones and Love (2011), Eberhardt and Danks (2011), Colombo and Seriès (2012), Marcus and Davis (2013), Marcus and Davis (2015), Block (2018), Colombo, Elkin, and Hartmann (2019), and Mandelbaum (2019).

[18]See Goodman, Frank, et al. (2015), Zednik and Jäkel (2016), Icard (2018), and Rescorla (2020).

[19]Though it should also be noted that it is not Bayesianism per se that will be doing the work. What will be important to my account is the use of generative random sampling in hypothesis testing. This can be separated from many of the claims Bayesians make—a generative random sampling algorithm for hypothesis testing need not approximate any kind of Bayesian norm—so they need not be right about all the important aspects of their

concepts *could* be learned, but how they *are* learned. This is not the place to adjudicate these debates. For now I will simply assume that the Bayesian approach, or some other approach relying on generative random sampling, is at least a serious option worth exploring for understanding how the human mind in fact works. If it can provide us with a counterexample to Fodor's premise (2), this would be a result of considerable interest.

3.2 The Monte Carlo Way

We've seen that Fodor's argument that concepts cannot be learned through hypothesis testing turns on an application of AIA. But AIA fails for just the kind of process that many cognitive scientists propose is our actual way of doing hypothesis testing. Putting these ideas together, we can show how Fodor's Circle can be avoided in principle, and may well be avoided in practice.

Just as Cindy actually drew the 58-ball without having an ability to draw the 58-ball, someone might actually think a thought without having the ability to think that thought. This might happen if thinking the thought is an unlikely result of a random process. So if an agent does hypothesis testing through an MCMC method, the fact that they actually sampled the hypothesis h out of the vast hypothesis space $\mathcal{H}$ is a very unlikely result of a random process. So it may well turn out that they did not have the ability to think h, even if they in fact thought it. And so they may have considered a hypothesis involving a concept C even if they do not have the ability to think C-thoughts. So their actual consideration of that hypothesis does not show that they already had the concept, even granting that having such an ability suffices for having the concept. And it may be that they acquire an ability to think C thoughts as a result of confirming h. So if we accept the cognitive scientists' proposal that concepts *are* acquired by

view, just this one component of it.

some random sampling-based method of hypothesis testing in learning concepts, then we'll have avoided Fodor's Circle. It is possible—and indeed plausible—that we learn concepts through hypothesis testing.

That's the Monte Carlo Way out of Fodor's Circle, stated relatively abstractly. It will be useful to see how it applies to a concrete theory of concept acquisition.

3.3 A Monte Carlo Way of Learning about Magnets

After making relatively few observations of some domain of objects, children generate and adopt theories of how objects in this domain work and why they work this way. Ullman, Goodman, and Tenenbaum (2012) give an account about how children discover and adopt these theories, and how they learn new concepts in the process.[20] To illustrate how the Monte Carlo Way works in practice, let's look at how it applies in the case of Ullman et al.'s account of how children learn theories of magnetism.

Ullman et al.'s account of theory learning has a few important parts. To begin with, there is assumed to be some *data* about the domain in question. This consists of some observations the child makes about, for example, which objects stick together or repel each other. The child is assumed to be trying to learn the correct *theory* of the domain in question. In Ullman et al.'s framework, theories are taken to be conjunctions of laws which specify conditions under which certain properties and relations hold. For example: a theory might have as a part the law that if F(object1) and F(object2), then R(object1, object2). Since a theory is just a statement of relational laws, to connect it to the world we also need a specification of which objects have which properties. In other words, children are also interested in which *model* of the correct theory is the right one.

To learn a theory, children start with a hypothesis space determined

[20] See Bonawitz et al. (2019) for a recent follow up offering more fine-grained empirical support for a model of this kind.

by a probabilistic rewrite grammar. Beginning with a start symbol, the grammar allows it to be transformed according to a variety of transition rules, each with a specified probability, eventually resulting in a string of terminal symbols that cannot be rewritten. In this case, the string of terminal symbols is a theory—a statement of a conjunction of laws. Each resulting theory will have a probability assigned, thus all the 'grammatical' hypotheses will have prior probabilities. Among the transformation rules will be rules which allow the introduction of arbitrarily many predicate and relation terms.

The posterior likelihood of a theory given some observed data is given by the probabilities of all the possible models of the theory and the how likely each model makes the observed data. So we have an infinite hypothesis space of theories, and the likelihood of any particular theory is given by what happens with a huge range of possible models satisfying that theory. This is just the kind of case calling for an approximation algorithm.

Following Goodman, Tenenbaum, et al. (2008), Ullman et al. propose a grammar-based MCMC algorithm that samples theories from the posterior distribution over theories conditioned on the data. To start, an initial theory from the hypothesis space is randomly generated. The likelihood of this theory given the observed data is estimated by a separate sampling algorithm over the models of the theory. Then the sampled theory is modified in a random way, by replacing a randomly chosen part of the theory with a randomly generated string from the grammar of theories. At this point, the likelihood of the new theory is estimated, and compared with the likelihood of the old theory. If the new theory makes the observed data more likely than the old theory, then it is accepted as the new current sample. If it doesn't, it is accepted as the current sample with probability proportional to the relative likelihoods of the data according to the two theories. This process is repeated many times, exploring a small part of the hypothesis space in a way that is random, yet biased towards theories that

do a better job predicting the observed data.

One way to use the results of this process would be estimate a posterior distribution over the hypothesis space, resulting in a probability distribution over the theory space, rather than a single accepted theory.[21] Ullman et al., however, treat the algorithm as a search algorithm, one trying to find the most likely theory. So after a certain number of iterations of the algorithm, the 'current' theory is treated as the accepted one, at least until new data comes in.[22]

As with many other MCMC algorithms attempting to find a global maximum—the most likely hypothesis—rather than the whole probability distribution, Ullman et al.'s algorithm involves *simulated annealing*, which over time reduces the chance that the learner will accept as a new sample a theory with lower probability than the current sample.[23] This means that the exploration will tend to stabilize in some local maximum of the hypothesis space, but only after the search has been going for some time.

Using such an algorithm, a child can learn and come to use theories containing predicates that do not already appear in any theory the child already knows or any beliefs they already have, since among the rewrite rules of the grammar that generates the hypotheses are rules for introducing arbitrary new predicates which can be constrained by laws of the theory. They may come to accept a theory that they had not been able to think prior to running the algorithm.

Consider, for example, a child who after some observations accepts a theory which groups objects into two classes: magnets, which are taken

[21] Which is one standard Bayesian approach to theory learning. See, for example, Thaker, Tenenbaum, and Gershman (2017).

[22] On one view, this algorithm will be re-run every time one makes any observation, or any observation involving the predicates of the theory in question. But one can imagine restricting this in various ways. Perhaps, e.g., the algorithm is only rerun when one makes a sufficiently surprising or puzzling observation.

[23] See Andrieu et al. (2003, pp. 18–20) for a discussion of simulated annealing in MCMC algorithms more generally.

to attract one another; and non-magnets, which neither attract each other nor are attracted by magnets. Suppose at this point they have never had any thoughts about objects which we would call magnetic but not magnets: objects which are attracted by magnets but which do not attract each other. After making some new observations and running the Ullman et al. algorithm again, the child might 'consider' a theory which postulates the property of being magnetic, as a random modification to the simpler theory they currently accept. Since such a theory will make the observed interactions more likely, it will be accepted as a replacement for simpler theories, and will be what they use to think about the objects they encounter, as well as their starting point for later runs of the algorithm.

According to Fodor's Circle, this cannot be a case of concept learning, since in order to think hypotheses about objects' being magnetic, the child must have already been able to think of objects' being magnetic, so must have already had the concept MAGNETIC. And this is where the Monte Carlo Way applies: the child does in fact think of an object's being magnetic while considering this hypothesis, but since AIA is false, it does not follow that they already had the ability to think of an object's being magnetic. Indeed, since in the first instance they only managed to think this thought as an unlikely result of this particular random 'toss' of theory generation, it is just the sort of case where we expect AIA to fail. We can suppose that at the time, nothing else they might have done or thought would have brought it about that they would think of objects as magnetic. So we should not say that the child already had the ability to think of objects as being magnetic at the time they considered the hypothesis in question. So even assuming AICP, we do not need to grant that they already possessed the concept.

So it is reasonable to hold that the child starts without the ability to think MAGNETIC-involving thoughts.[24] And they also end up with such an ability

[24]Perhaps one might say the child already thought of things being magnetic, but confused this with their being a magnet. For approaches to modelling this kind of confusion, see Ripley (2018) and works cited therein. Whether something like this is the right approach

as a result of doing hypothesis testing. As part of a search algorithm, they consider a hypothesis with MAGNETIC (still without having the ability to think such thoughts), which is confirmed relative to their previous theories by the data. Because of this, this hypothesis will be accepted as the new sample, and so will be the base for further exploration of the hypothesis space. This will make it highly likely that the next theories to be considered will also involve MAGNETIC. And if, as is plausible, the theory that is in accepted at the end of the search involves it, it will be available to be used by the child to understand and make predictions about their environment. So as a result of the confirmation of the first theory with MAGNETIC they considered, the child will come to have the ability to think thoughts with MAGNETIC, and by AICP, must have the concept by this point. And given that they come to have this ability through the hypothesis being confirmed, it seems that they have in fact learned MAGNETISM through hypothesis testing.

Concept learning by an AIA-violating kind of hypothesis testing is a plausible theory of how we actually come to have many of our concepts. The Monte Carlo Way successfully gets us out of Fodor's Circle.

4. Obstacles Along the Way

There are several important objections to the Monte Carlo Way. I will now consider and answer some of the most pressing ones.

4.1 Rationality with Randomness?

Even if we grant that MCMC algorithms can count as hypothesis testing, we might worry that the kind of randomness involved keeps them from

in this case will turn on delicate issues about concept individuation which I concede might go either way. But this is just a special feature of the magnet case that won't be present in general. Some additions to theories will not be closely enough related to the previous theories to reasonably count as modifications as opposed to addition of concepts.

counting as rational processes. If we came to have a concept as a result of a random process, how can this count as acquiring it in a rationally evaluable way, as opposed to a matter of dumb luck? So if we're going along with Fodor's claim that a process must be rationally evaluable to count as learning, doesn't this method of hypothesis testing fail to count as learning? In which case, we can resuscitate Fodor's argument by replacing 'hypothesis testing' with 'rationally evaluable hypothesis testing'.

In reply, we should observe that even if the random sampling itself cannot be a rational process, this does not mean the process leading to the acquisition of the concept is non-rational. Not every part of a rational process needs itself to be a rational process. And there are many rational processes that involve randomness in significant ways (see Icard (2019)).

It is also important to note that on the picture presented in §3.3, MAGNETIC is acquired not just because it is part of a theory that is randomly sampled, but because such a theory is *confirmed by the data* relative to the previously sampled theory. Had it been disconfirmed, the algorithm would have probably returned to the previous theory, rather than stabilizing in the region of theories that include MAGNETISM. The proposed learning process, then, crucially involves a stage of epistemic evaluation. So it is not *only* as a result of random sample being generated that one acquires the concept.

These points, I think, are enough to answer the objection. A fully satisfying defense of this process being a rational one would require us to defend a theory of what it takes for any process at all to be rational, and show that it applies in this case. But this is a task for another time. For now I will just say that the randomness involved does not give us good reason to think that this cannot be done.

4.2 Representing $\mathcal{H}$ without Representing h?

Here's another worry: doesn't the learner using an algorithm like Ullman et al.'s need to be able to represent the hypothesis space $\mathcal{H}$ from the

beginning? And wouldn't this require them to already be able to represent each hypothesis $h \in \mathcal{H}$? But then the learner would need to have the concepts required for being able to think any of the theories they could come to have, be it MAGNETIC or CARBURETOR, which lands us right back in Fodor's conclusion: this cannot be a way of learning concepts, since it requires that one possess the concepts already.[25]

First, it is not clear that the learner needs to represent $\mathcal{H}$. They just need to be able to generate hypotheses in accordance with the probabilistic rewrite grammar that determines $\mathcal{H}$ and the learner's prior's over it. Even granting that this requires some explicit representation of the grammar's transition rules and their probabilities, it is not clear that this should suffice for holding that the agent is thereby representing $\mathcal{H}$ as a whole.

Second, even if we grant that this would count as representing $\mathcal{H}$, it does not follow that this must count as representing each $h \in \mathcal{H}$. One might 'cognize' a natural language without representing each grammatical sentence in the language, perceive the speckled hen without representing each particular speckle, and think of the set of real numbers without each real number being such one represents that number in particular. So even if we say that the learner must be able to represent $\mathcal{H}$ in order to learn via some Bayesian approximation algorithm, this does not mean they already

[25] Indeed, Ullman et al. themselves seem to concede as much:

> There is a sense in which, at the computational level, the learner already must begin the learning process with all the laws and concepts needed to represent a theory already accessible. Otherwise the necessary hypothesis spaces and could not be defined. In this sense, Fodor's skepticism on the prospects for learning or constructing truly novel concepts is justified. Learning cannot really involve the discovery of anything "new", but merely the changing of one's degree of belief in a theory, transporting probability mass from one part of the hypothesis space onto another.
>
> Ullman, Goodman, and Tenenbaum (2012, p. 478)

And though they immediately go on to say that there is also a sense in which learners do make genuine discoveries of new concepts and laws, I think they've already conceded to much to Fodor here.

need to be able to represent each of the theories which they might come to consider through running the algorithm.

Nevertheless, there remains a concern. The learning algorithms in question don't require just any representation of $\mathcal{H}$, but one which allows for sampling its members. But how could one sample from it if the individual potential samples aren't already there? For Cindy to draw the 58-ball out of the bag, the 58-ball must already be in the bag. And it seems the only way the members of $\mathcal{H}$ could already be there to be sampled would be by their each being represented. So it looks like the learner using an MCMC algorithm does need to be able to represent each hypothesis before they come to consider it.

This concern, however, turns on pushing the balls-from-a-bag picture of sampling beyond its appropriate limits. The approximation algorithms in question, and the Ullman et al. algorithm in particular, rely on *generative* random sampling. The samples are not sitting in a mental bag to be drawn from when required. Rather, they are constructed on the fly according to a partly randomized procedure. In the Ullman et al. algorithm, what's required is some procedure which is able to generate theories specified by the hypothesis grammar with the probabilities determined by those given for the grammar's transition rules, as well as a procedure for taking a previously generated output from the grammar, randomly picking a place to modify it, and generating a new theory in $\mathcal{H}$ on that basis. None of this requires that the theories in $\mathcal{H}$ are already represented by the learner prior to running the learning algorithm any more than being able to generate some sentence in English, or some sequence of legal chess moves, requires that you already had a representation of that sentence or that sequence of chess moves.

Nor does being able to generate any member of $\mathcal{H}$ imply that for any member of $\mathcal{H}$, one can generate it. It's not just AIA that fails for this kind of randomized process, but also distribution of ability over disjunction,

as Kenny (1975, p. 137) also pointed out. Cindy has the ability to draw a ball from the bag—i.e., has the ability to [draw the 1-ball $\vee$ draw the 2-ball $\vee \ldots \vee$ draw the 10,000-ball]—this does not mean she has the ability to draw the 1-ball $\wedge$ the ability to draw the 2-ball $\wedge \ldots \wedge$ the ability to draw the 10,000-ball. What's required for the relevant sampling algorithms are abilities to generate samples from $\mathcal{H}$, but these do not imply that one already has abilities to think each member of $\mathcal{H}$. So one need not have the concepts required for thinking each member of $\mathcal{H}$ in order to run such algorithms.

4.3 Evaluating h?

It is important to the proposal we're considering that the learner not only think a given hypothesis h, but also evaluate it. The proposal is one about how concepts are learned through hypothesis *testing*, not mere hypothesis generation. An objector might allow that coming up with h in the first place could be the result of a random process that doesn't require an ability to think h, but deny that the evaluation stage could be performed through luck or randomness, and so take it to require an ability to think h. On the Ullman et al. account, evaluation will involve estimating the likelihood of h by generating many models satisfying the theory it expresses and checking whether the observed data obtain in those models. This use of h seems systematic, not like a matter of luck. So mustn't the learner already be able to think h after all, at least at the point at which it is being evaluated? Circularity threatens once again.

To avoid the threat, we should distinguish between being able to ϕ and being able to ϕ from a position partly along the way to ϕ-ing. The latter need not entail the former. To see this, consider the following variant of Lucky Hole-in-one.

Lucky Birdie

> Bob is a terrible golfer, often badly messing up even easy shots. On this occasion, though, he luckily swings with decent form, then a lucky bounce off a tree and unexpected gust of wind carries his ball onto the green within a foot of the hole. He then makes the easy putt for a birdie. Though Bob did in fact birdie, he did not have the ability to birdie.

Through luck, Bob is put in the position far along the way to getting a birdie. He does have the ability to follow through and birdie *from this position*, as a matter of skill rather than luck. However, at no point does he have the ability to simply birdie the hole. In particular, we should not think that once the ball has landed on the green Bob has suddenly acquired the ability to birdie the hole. He continues to lack this ability, as the lucky darts player continues to lack the ability to hit the bullseye in the millisecond before their dart in fact hits it. If Bob were to play the hole again, he would almost certainly fail to birdie.

We can say something similar about the learner's evaluation of h. Partly through luck, the learner comes to consider h. They may then, from this position, be able to follow through and successfully test h by generating models of it and so on. Like Bob's putting, this follow-through may be a matter of skill rather than luck, and may be quite systematic. However, at no point does the learner need to have the ability to simply think or test h. In particular, we should not think that once h has been generated, a learner originally unable to think or test h has suddenly acquired the ability to do so. They acquire such abilities, on this proposal, only once h has been sufficiently confirmed and the learning algorithm stabilized on it through simulated annealing. Until then, if faced with with similar data or other good occasions for thinking or testing h, the learner would almost certainly fail to do so.

If this reply to the objection is on the right track, we will want our theory of concept learning to tell us about the details of the postulated ability to

evaluate a wide range of generated hypotheses. This will likely be bound up with details about how the generated hypotheses are in fact represented, another issue that deserves a more thorough treatment than can be given here. And it should be emphasized that it is an empirical question whether in human learning the generation of some hypotheses depends on luck until they have been sufficiently confirmed. One can imagine learners who acquire the ability to rethink some thought as soon as they've generated it once, regardless of whether it has been confirmed or disconfirmed. We should thus not take the reply here to be the end of the theoretical story. It aims only to show how we can coherently hold that a learner might evaluate h in the systematic way outlined by Ullman et al. without taking the learner to already have an ability to think h.

4.4 What Kind of Ability?

Distinguishing kinds of ability has a venerable history in philosophy. So when applying a claim about ability to some particular philosophical issue, one must take care to avoid a bait-and-switch, and check that the kind of ability that the claim is true of is the right kind of ability for the philosophical application at hand. One might worry, then, that even if AIA fails for certain kinds of abilities, it does hold for the *relevant* kinds of abilities, in which case the Monte Carlo Way looks like a dead end. That is, if AIA fails for the kinds of abilities that concept possession entails, then I have not found a good way of responding to Fodor's argument against the possibility of concept learning by hypothesis testing.

To turn this concern into a proper objection, we would need to identify a kind of ability for which AIA is valid, then plausibly claim that this sense of ability suffices for concept possession, as AICP requires. But this is by no means an easy task.

It is plausible that 'able to' is ambiguous in English, and that one of the senses does indeed validate AIA. Bhatt (1999, Ch. 5) observes that 'was able

to' has different implications when it is read as an episodic way, as in (2-a), and when it is read in a generic way, as in (2-b).

(2) a. Yesterday, John was able to eat five apples in an hour.
b. In those days, John was able to eat five apples in an hour.

AIA fails for the generic reading, for processes that turn on luck or randomness in a significant enough way. On its episodic reading, however, 'able to' is basically equivalent to 'managed to'. And it is at least plausible that actuality suffices for possessing ability in this sense, even when the result was a matter of luck: perhaps yesterday, Bob was able (managed) to hit a hole in one, Cindy was able (managed) to draw the 58-ball, and Kenny's darts player was able (managed) to hit the bullseye.

The problem for turning this into an objection is that on this sense of 'able to', ability is not just sufficient, but necessary. If someone didn't ϕ, they didn't manage to ϕ, and so were not, in the relevant sense, able to ϕ. And no serious view of concept possession can require one to be actually thinking of bears in order to have the concept BEAR. So there is no reason to take this to be the kind of ability closely connected with concept possession.

As far as I know, there is no good reason to take the ordinary English sense of 'able to' to have any other readings. So at least if the objector is looking for a suitable sense of 'ability' in English, I do not think their search will succeed. Perhaps, though, no ordinary sense of 'able to' is suitable for serious theorizing about the mind, and so we need to stipulate a technical sense of 'is able to' instead. And perhaps the right technical sense of 'able to' will be one on which AIA holds. So there is still hope that this objection will block the Monte Carlo Way out of Fodor's Circle.

Well, perhaps. Fodor himself did not specify any technical sense of 'ability', but this doesn't mean it can't be done. The trick would be to specify some sense of ability for which AIA fails but AICP still holds (whether self-evidently or not). I cannot rule out that someone could find such a

sense, but I am skeptical. In order to validate AIA, the sense of ability will need to be a fairly weak one. But why should we care about such a weak sense of ability for theorizing about concept possession? My own suspicion is that the notion of ability will need to be precisified for certain theoretical purposes, but that what we're likely to find is that it is still a relatively robust form of ability—one for which AIA fails—that we take to entail concept possession.

5. Conclusion

Since AIA is false, there is a gap in Fodor's argument against the possibility of concept learning by hypothesis testing. Models of learning that rely on generative random sampling sail neatly through this gap. Concepts can be learned through hypothesis testing, and it is plausible that this is how humans learn at least some of their concepts. We can escape Fodor's Circle by taking the Monte Carlo Way.

References

Andrieu, Cristophe et al. (2003). "An Introduction to MCMC for Machine Learning". In: *Machine Learning* 50, pp. 5–43.

Austin, J. L. (1956). "Ifs and Cans". In: *Proceedings of the British Academy* 42, pp. 109–132. Reprinted in Austin (1961, pp. 153–180).

— (1961). *Philosophical Papers*. Oxford University Press.

Bhatt, Rajesh (1999). "Covert Modality in Non-finite Contexts". PhD thesis. University of Pennsylvania.

Block, Ned (2018). "If perception is probabilistic, why does it not seem probabilistic?" In: *Philosophical Transactions of the Royal Society B* 373.

Bonawitz, Elizabeth et al. (2019). "Sticking to the Evidence? A Behavioral and Computational Case Study of Micro-Theory Change in the Domain of Magnetism". In: *Cognitive Science* 43.

Buijsman, Stefan (2019). "Acquiring mathematical concepts: The viability of hypothesis testing". In: *Mind and Language*.

Carey, Susan (2009). *The Origin of Concepts*. Oxford University Press.

— (2011). "Concept innateness, concept continuity, and bootstrapping". In: *The Behavioral and Brain Sciences* 34.3, pp. 152–167.

— (2015). "Why Theories of Concepts Should Not Ignore the Problem of Acquisition". In: *The Conceptual Mind: New Directions in the Study of Concepts*. Ed. by Eric Margolis and Stephen Laurence. MIT Press, pp. 415–454.

Clark, Andy (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Colombo, Matteo, Lee Elkin, and Stephan Hartmann (2019). "Being Realist about Bayes, and the Predictive Processing Theory of Mind". In: *British Journal of Philosophy of Science*.

Colombo, Matteo and Peggy Seriès (2012). "Bayes in the Brain—On Bayesian Modelling in Neuroscience". In: *British Journal of Philosophy of Science* 63, pp. 697–723.

Denison, Stephanie et al. (2013). "Rational variability in children's causal inferences: The Sampling Hypothesis". In: *Cognition* 126, pp. 285–300.

Earman, John (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. The MIT Press.

Easwaran, Kenny (2011a). "Bayesianism I: Introduction and Arguments in Favor". In: *Philosophy Compass* 6.5, pp. 312–320.

— (2011b). "Bayesianism II: Applications and Criticisms". In: *Philosophy Compass* 6.5, pp. 321–332.

Eberhardt, Frederick and David Danks (2011). "Confirmation in the Cognitive Sciences: The Problematic Case of Bayesian Models". In: *Minds & Machines* 21, pp. 389–410.

Fine, Gail (2014). *The Possibility of Inquiry: Meno's Paradox from Socrates to Sextus*. Oxford University Press.

Fodor, Jerry A. (1975). *The Language of Thought*. New York: Thomas Y. Crowell Company.

— (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford Cognitive Science Series. Oxford University Press.

— (2008). *LOT 2: The Language of Thought Revisited*. Oxford University Press.

Friston, Karl (2010). "The free-energy principle: a unified brain theory?" In: *Nature Reviews: Neuroscience* 11, pp. 127–138.

Goodman, Noah D., Michael C. Frank, et al. (2015). "Relevant and Robust: A Response to Marcus and Davis (2013)". In: *Psychological Science* 26.4, pp. 539–541.

Goodman, Noah D., Joshua B. Tenenbaum, et al. (2008). "A Rational Analysis of Rule-Based Concept Learning". In: *Cognitive Science* 32, pp. 108–154.

Griffiths, Thomas L., Charles Kemp, and Joshua B. Tenenbaum (2008). "Bayesian models of cognition". In: *The Cambridge Handbook of Com-*

putational Psychology. Ed. by Ron Sun. Cambridge University Press, pp. 59–100.

Griffiths, Thomas L., Edward Vul, and Adam N. Sanborn (2012). "Bridging Levels of Analysis for Probabilistic Models of Cognition". In: *Current Directions in Psychological Science* 21.4, pp. 263–268.

Hilpinen, Risto (1969). "An Analysis of Relativised Modalities". In: *Philosophical Logic*. Ed. by J. W. Davis, D. J. Hockney, and W. K. Wilson. D. Reidel Publishing Company.

Hohwy, Jakob (2013). *The Predictive Mind*. Oxford University Press.

Icard, Thomas F. (2016). "Subjective Probability as Sampling Propensity". In: *Review of Philosophy and Psychology* 7, pp. 863–903.

— (2018). "Bayes, bounds, and rational analysis". In: *Philosophy of Science* 85.1, pp. 79–101.

— (2019). "Why Be Random?" In: *Mind*.

Jones, Matt and Bradley C. Love (2011). "Bayesian Fundamentalism or Enlightenment? On the explantory status and theoretical contributions of Bayesian models of cognition". In: *Behavioral and Brain Sciences* 34, pp. 169–231.

Kenny, Anthony (1975). *Will, Freedom, and Power*. Oxford: Blackwell.

Kersten, Daniel, Pascal Mamassian, and Alan Yuille (2004). "Object Perception as Bayesian Inference". In: *Annual Review of Psychology* 55, pp. 271–304.

Körding, Konrad P. and Daniel M. Wolpert (2004). "Bayesian integration in sensorimotor learning". In: *Nature* 427, pp. 244–247.

Kratzer, Angelika (1977). "What *Must* and *Can* Must and Can Mean". In: *Linguistics and Philosophy* 1, pp. 337–355. Reprinted with revisions in Kratzer (2012, pp. 1–20).

— (1981). "The Notional Category of Modality". In: *Words, Worlds, and Contexts*. Ed. by H. J. Eikmeyer and H. Rieser. de Gruyter, pp. 38–74. Reprinted with revisions in Kratzer (2012, pp. 21–69).

Kratzer, Angelika (2012). *Modals and Conditionals*. Oxford University Press.

Kruschke, John K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press.

Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum (2015). "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266, pp. 1332–1338.

Lake, Brenden M., Tomer D. Ullman, et al. (2017). "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40.

Lee, Michael D. and Eric-Jan Wagenmakers (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lloyd, Kevin et al. (2019). "Why Higher Working Memory Capacity May Help You Learn: Sampling, Search, and Degrees of Approximation". In: *Cognitive Science* 43.

MacKay, David J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Maier, John (2015). "The Agentive Modalities". In: *Philosophy and Phenomenological Research* 90.1, pp. 113–134.

Mandelbaum, Eric (2019). "Troubles with Bayesianism: An introduction to the psychological immune system". In: *Mind & Language* 34, pp. 141–157.

Mandelkern, Matthew, Ginger Schultheis, and David Boylan (2017). "Agentive Modals". In: *Philosophical Review* 126.3, pp. 301–343.

Marcus, Gary F. and Ernest Davis (2013). "How Robust Are Probabilistic Models of Higher-Level Cognition?" In: *Psychological Science* 24.12, pp. 2351–2360.

— (2015). "Still Searching for Principles: A Response to Goodman et al. (2015)". In: *Psychological Science* 26.4, pp. 542–544.

Margolis, Eric and Stephen Laurence (2011). "Learning Matters: The Role of Learning in Concept Acquisition". In: *Mind & Language* 26.5, pp. 507–539.

Mele, Alfred R. (2003). "Agents' Abilities". In: *Noûs* 37.3, pp. 447–470.

Ong, Desmond C., Jamil Zaki, and Noah D. Goodman (2015). "Affective cognition: Exploring lay theories of emotion". In: *Cognition* 143, pp. 141–162.

Perfors, Amy, Joshua B. Tenenbaum, Thomas L. Griffiths, et al. (2011). "A tutorial introduction to Bayesian models of cognitive development". In: *Cognition* 120, pp. 302–321.

Perfors, Amy, Joshua B. Tenenbaum, and Terry Regier (2011). "The learnability of abstract syntactic principles". In: *Cognition* 118, pp. 306–338.

Rescorla, Michael (2015). "Bayesian Perceptional Psychology". In: *The Oxford Handbook of the Philosophy of Perception*. Ed. by Mohan Matthen. Oxford University Press, pp. 694–716.

— (2016). "Bayesian Sensorimotor Psychology". In: *Mind & Language* 31.1, pp. 3–36.

— (2020). "A Realist Perspective on Bayesian Cognitive Science". In: *Inference and Consciousness*. Ed. by Timothy Chan and Anders Nes. Routledge.

Ripley, David (2018). "Blurring: An Approach to Conflation". In: *Notre Dame Journal of Formal Logic* 59.2, pp. 171–188.

Sanborn, Adam N. (2017). "Types of approximation for probabilistic cognition: Sampling and variational". In: *Brain and Cognition* 112, pp. 98–101.

Sanborn, Adam N. and Nick Chater (2016). "Bayesian Brains without Probabilities". In: *Trends in Cognitive Sciences* 20.12, pp. 883–893.

Schwarz, Wolfgang (2020). "Ability and Possibility". In: *Philosophers' Imprint* 20.6, pp. 1–21.

Spencer, Jack (2017). "Able to Do the Impossible". In: *Mind* 126.502, pp. 465–497.

Sprenger, Jan and Stephan Hartmann (2019). *Bayesian Philosophy of Science: Variations on a Theme by the Reverend Thomas Bayes*. Oxford University Press.

Sterelny, Kim (1989). "Fodor's Nativism". In: *Philosophical Studies* 55, pp. 119–141.

Sundström, Pär (2019). "Are Sensory Concepts Learned by "Abstraction" from Experience?" In: *Erkenntnis* 84, pp. 1159–1178.

Tenenbaum, Joshua B. (1999). "A Bayesian Framework for Concept Learning". PhD thesis. Massachusetts Institute of Technology.

Tenenbaum, Joshua B. et al. (2011). "How to Grow a Mind: Statistics, Structure, and Abstraction". In: *Science* 331.6022, pp. 1279–1285.

Thaker, Pratiksha, Joshua B. Tenenbaum, and Samuel J. Gershman (2017). "Online learning of symbolic concepts". In: *Journal of Mathematical Psychology* 77, pp. 10–20.

Ullman, Tomer D., Noah D. Goodman, and Joshua B. Tenenbaum (2012). "Theory learning as stochastic search in the language of thought". In: *Cognitive Development* 27, pp. 455–480.

Vihvelin, Kadri (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. Oxford University Press.

Vul, Edward et al. (2014). "One and Done? Optimal Decisions From Very Few Samples". In: *Cognitive Science* 28, pp. 599–637.

Weiskopf, Daniel A. (2008). "The origins of concepts". In: *Philosophical Studies* 140, pp. 359–384.

Xu, Fei and Joshua B. Tenenbaum (2007). "Word Learning as Bayesian Inference". In: *Psychological Review* 114.2, pp. 245–272.

Zednik, Carlos and Frank Jäkel (2016). "Bayesian reverse-engineering considered as a research strategy for cognitive science". In: *Synthese* 193, pp. 3951–3985.