Local search and the evolution of world models

Neil R. Bramley[1], Bonan Zhao[1], Tadeg Quillien[2], and Christopher G. Lucas[2]

[1] Department of Psychology

University of Edinburgh

[2] Institute of Language

Cognition & Computation

Informatics University of Edinburgh

Author Note

Correspondence address: Neil Bramley (neil.bramley@ed.ac.uk), Department of Psychology, Room S2, 7 George Square, University of Edinburgh EH8 9JZ, Scotland.

Acknowledgments: This work was supported by by an EPSRC New Investigator Grant (EP/T033967/1) to Bramley and Lucas. Many thanks to Aba Szollosi for his ideas and contributions to the first draft of this paper. Thanks to Stephanie Droop, Nick Chater and an anonymous reviewer for their constructive comments. This paper is dedicated to the ideas and legacy of an arch universal Darwinist, Douglas Alastair Bramley (1988-2022).

Keywords: learning; inference; concepts; search; evolution; approximation; MCMC; bootstrapping; adaptor grammar

Abstract

An open question regarding how people develop their models of the world is how new candidates are generated for consideration out of infinitely many possibilities. We discuss the role that evolutionary mechanisms play in this process. Specifically, we argue that when it comes to developing a global world model, innovation is necessarily incremental, involving the generation and selection among random local mutations and recombinations of (parts of) one's current model. We argue that, by narrowing and guiding exploration, this feature of cognitive search is what allows human learners to discover better theories, without ever grappling directly with the problem of finding a "global optimum", or best possible world model. We suggest this aspect of cognitive processing works analogously to how blind variation and selection mechanisms drive biological evolution. We propose algorithms developed for program synthesis provide candidate mechanisms for how human minds might achieve this. We discuss objections and implications of this perspective, finally suggesting that a better process-level understanding of how humans incrementally explore compositional theory spaces can shed light on how we think, and provide explanatory traction on fundamental cognitive biases including anchoring, probability matching and confirmation bias.

Keywords: learning; inference; concepts; search; evolution; approximation; MCMC; bootstrapping; adaptor grammar

Local search and the evolution of world models

Introduction

In an 1897 address to the British Association for the Advancement of Science Lord Kelvin claimed: *"There is nothing new to be discovered in physics now. All that remains is more and more precise measurement."* Later the same year, Rutherford discovered the electron, a peculiar particle that behaved like a wave in contradiction to prevailing physical theories of the day (Rutherford, 1911). This spurred a new generation of progress in physics leading to radically different formalisms like quantum theory that have unlocked a range of other insights and technologies.

The history of ideas is strewn with these "false summits" and moments of hubristic overconfidence by the victors of the day. Our most celebrated scientific and cultural innovations have almost all been usurped in our ongoing struggle to understand the world. In a similar way, the phenomenology of individual development has its own progression of "aha!" moments, in which one seems to land on new ways of thinking, or better solutions to problems that have puzzled us (Kounios & Beeman, 2009). Like the history of science, with the benefit of hindsight the opinions of our younger selves will often strike us as excruciatingly naive. With these general phenomena in mind, it is perhaps surprising that, as a field, we frequently seek to explain the products of cognition as approximately rational or optimal solutions to challenges faced by cognizers. Bayesian accounts take the mind to be engaged in (approximate) probabilistic inference about the nature of the environment (Chater, Oaksford, Hahn, & Heit, 2010; Griffiths & Tenenbaum, 2009; Howson & Urbach, 2006; Tenenbaum, Griffiths, & Kemp, 2006), taking actions that subserve this (Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018; Bramley, Lagnado, & Speekenbrink, 2015; Gureckis & Markant, 2012; Settles, 2009) and decisions that marginalize appropriately over our well-calibrated subjective uncertainty (Maloney & Mamassian, 2009; Oaksford & Chater, 2007). The closely related predictive processing tradition and "free energy principle" models the mind as minimizing its free energy, equivalent to maximizing model

probability (Gershman, 2019b), with this determining both our beliefs (Friston, 2010) and actions (Friston et al., 2016).

On the face of it, neither framework offers a direct solution to the question of how a cognizer's overall model space—that is, the set of concrete models or theories they optimize over—is formed or adapted; nor tells us how close someone's current theory is to perfection. At its core, Bayesian inference is a process of pure selection—all the hypotheses are already there for consideration and one simply updates their probabilities with evidence. Similarly, the free energy principle describes a process of model optimization—by adjusting parameters to minimize expected surprise, such an account is guaranteed to optimize its interactions with its environment with respect to the representational expressivity of the variational approximation it starts with. Both frameworks seem to leave unanswered questions as to where the state space of possible models originates from, and if and how it can be extended. In this paper, we describe recent algorithmic approaches for approximating intractable probabilistic inference problems and argue they help address these questions. Moreover we reframe the issue in evolutionary terms. We argue for a parallel between biological evolution and the haphazard growth and revision of an individual's system of concepts, global theory, or world model. In doing so, we highlight how the algorithms we describe fundamentally operate via blind variation and selective retention. In particular, we argue that the evolutionary perspective reveals why a mechanism for producing random (but local) variation (D. T. Campbell, 1960) is both a core limitation of *how* minds work and a core feature of *why* minds work.

Our thesis

We propose that the structure of the conceptual system represented within an individual's mind is discovered and adapted through local search over an open pool of candidates, with small incremental changes (randomly produced, selectively retained) driving innovations just as natural variation does in biological evolution. The central idea of evolutionary theory is that the conditions needed for any organized system to emerge is

the existence of some mechanism of repeated *blind variation* paired with some mechanism and vehicle for *selective retention* (Darwin, 1859/2004; Dawkins, 1983). Indeed, this is the only type of mechanism yet discovered that we know to be capable of constructing complex functional design in the absence of a designer (Dawkins, 1986). Universal Darwinism generalizes this idea beyond biological kinds to suggest analogous mechanisms of selection explain the emergence of all forms of functional complexity, including any conceptual system, model or theory within our minds (J. O. Campbell, 2016; Hodgson, 2005; Popper et al., 1979; Sydow, 2012).

The notion that evolutionary mechanisms describe aspects of cognitive processing has been raised a handful of times (cf. Ashby, 1952; Bourgin, Abbott, Griffiths, Smith, & Vul, 2014; D. T. Campbell, 1960; Pinker, 2003; Plotkin, 1997; Suchow, Bourgin, & Griffiths, 2017). In particular, advancing this kind of view in psychology, D. T. Campbell (1960) argues that *any* cognitive process that appears to do something smarter or more goal-driven than blind variation and selective retention, can only be doing so because it is exploiting earlier established expectations about the domain within which it is operating. To illustrate this, Campbell takes visual perception as a paradigmatic example of the kind of cognitive apparatus we do not think of as blind. He describes how individual visual receptors can be thought of as exploring the possibilities of locomotion in many different directions, with this based on an earlier discovery that the measuring of arriving photons is an effective substitute for a more primitive and energetically costly process of random locomotion: *"For the 'blindness' of an eyeless animal there has been substituted a process so efficient that we use it naively as a model for direct, unmediated knowing."* (D. T. Campbell, 1960, p383). The key point is that the visual system—and, by the same argument, the other inductive biases, heuristics and tricks we bring to the table—can be framed as inductive shortcuts that were discovered, refined and improved via a long chain of earlier serendipitous discoveries that were, each in turn, selected for. In other words, the results of our trial-and-error processes become genetically encoded (on a phylogenetic

timescale) or learned (on an ontogenetic timescale). Whenever they are applied in novel contexts, we are exploiting the results of these processes—and so, in a sense, relying on the results of earlier blind processes.

Putting blind variation at the heart of conceptual change may seem like a skeptical thesis, undercutting the prospect of fully understanding the mind, not to mention limiting its potential for agency. However, we believe some degree of blind variation may be an essential design principle for any system to be capable of inductive reasoning, and moreover that this has implications for how to understand idea generation and distinguish it from the more pedestrian forms of inference we engage in.

Campbell's core claim is that evolutionary mechanisms determine how minds grow and adapt their content: That is, they shape how we form our concepts, theories and models of the world. We will argue on these lines that it is through an accumulation of incremental changes (randomly produced, selectively retained) that new conceptual structure is first discovered, meaning "randomness" plays a starring role in concept change in the same way natural variation does in biological evolution.[1] A consequence of this idea is that it is only among already-developed candidates that prior expectations can form, or deliberate (model based) optimization (or selection) can occur. Of course, a large part of everyday cognition is about employing one's existing world model to make on-the-spot predictions, explanations, plans and so on. Moreover, the new ideas we generate and select among are deeply connected to our earlier ideas. However, we are suggesting that genuine conceptual innovations are exactly the point at which this foresight falls away. The fact that our new ideas are shaped, constrained and evaluated against our existing ideas is the "locality" we refer to in the title of the paper and that is embodied by the algorithms we describe in the subsequent sections.

Just as modern evolutionary theory is very specific about where variation and

[1] We use random, in the subjective sense that whatever mechanism induces the variation is exogenous to, and unpredictable by, the mechanism that is using it.

selection manifest in biological evolution, we think cognitive scientists can and should be specific about how variation and selection manifest in cognition. Moreover, just as biological evolution is characterized by its unpredictability, incrementality, and path dependence, we think that cognitive scientists should consider adopting the idea that the mechanisms of concept change: (1) Rely to some extent on random variation, yet (2) Put tight constraints on what new concepts are within reach of a cognizer at any moment. Several extant algorithms for nonparametric approximation to probabilistic inference provide candidate evolutionary mechanisms in the sense that they work by mirroring established mechanisms of biological evolution. We will suggest that taking an evolutionary view of the role and limitations of these mechanisms helps defuse persistent tensions in the learning and decision making literature, for example providing a parsimonious explanation for why we exhibit anchoring, probability matching and confirmation bias.

To develop this argument, we first sketch some of the ways that blind variation and selective retention operate in biological evolution. We then relate these to a "learning as program induction" framework (Bramley & Xu, 2023; Chater & Oaksford, 2013; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, 2021; Rule, Tenenbaum, & Piantadosi, 2020; Ullman, Goodman, & Tenenbaum, 2012; Zhao, Bramley, & Lucas, 2022; Zhao, Lucas, & Bramley, 2022, 2023). We argue this framework captures the latent possibility space and the mechanisms needed for concepts to evolve locally and incrementally within an individual mind. We discuss algorithms developed for program induction and relate them to evolutionary principles of blind variation and selection. In particular, we highlight how tree-mutating *Markov Chain Monte Carlo* methods can capture how a cognizer can tinker with and improve on their current model by generating random hypothetical mutations and selectively retaining them. We then discuss how integrating combinatory logic with conceptual bootstrapping in an *Adaptor Grammar* scheme can additionally capture how a cognizer's repertoire of future mutations or conceptual moves can grow alongside their concepts, through the selective caching of promising conceptual fragments. We will also

argue that this kind of framework offers insight about the origin of hierarchical structure and abstraction in our conceptual systems. We draw a soft analogy between these algorithms and familiar mechanisms of biological evolution and conclude by suggesting that they capture something fundamental about how minds operate that is not obvious from the perspective of the computational level theories they are more normally used to approximate.

Established mechanisms of biological evolution

The now-standard explanation for the presence of complex functional forms in biology is the operation of natural selection (Darwin, 1859/2004; Williams, 1966). Natural selection combines a process of blind variation with a process of selective retention. Random genetic mutations create variation in the pool of genes. The combination of genes that are more successful at spreading themselves through the population (typically by conferring advantages to the organism, but sometimes disadvantages) are then selectively retained through the greater survival and reproduction prospects of organisms containing that combination (Dawkins, 1976). Unfolding over many generations, this process "designs" organisms that are good at replicating their genes by solving the adaptive challenges in their environmental niche (Williams, 1966).

The process of evolution by natural selection is "blind" in the sense that it cannot plan ahead (Dawkins, 1986). Genetic mutation is random: It is not biased toward new genetic variants that are systematically better. Selection then favors the genetic variants that turn out better at contributing to their own replication, but it cannot anticipate whether a given variant might eventually be useful at a later stage of evolution. For example, the early ancestor of the eye was probably a simple layer of photosensitive cells that could detect the direction of the light. This simple design was successful because it increased the reproductive success of the organism, not because evolution could anticipate that fully-realized eyes would be useful millions of years later (Dawkins, 1986; Dennett, 1995; Jacob, 1977).

Evolution by natural selection is a process of optimization by local search—in the sense that nearby possibilities are evaluated and the more successful ones adopted leading to a tendency to climb "uphill" in fitness space (J. O. Campbell, 2016). In principle, a swathe of mutations could align in a single generation and directly create a new "good trick" or even a completely different organism. However, this is possible only in the same sense that a monkey hitting 5 million random keys on a typewriter could reproduce the complete works of Shakespeare (Borges, 1941/1998). In practice, biologically successful arrangements of matter as complex and structured as humans are astronomically unlikely to arise from random coincidence: One would not expect anything interesting to come from arbitrarily arranging the 7 octillion atoms that make up a human body. Once a nontrivial organism has evolved, the larger the random leap it takes from its working design, the greater the chance that it will land on something catastrophically worse. For this reason, as far as we know, all of the complex forms we actually encounter in biology are those that were able to evolve through a long series of tiny incremental changes, that each improved, or at least did not substantially harm, the fitness of the organism (Carroll, 2005; Dawkins, 1986; Jacob, 1977). For such changes to accumulate into a highly complex functional form, the units of selection (in the context of evolution, the genes), have to persist long enough (via reproduction on the level of the organism), or copy themselves accurately enough, that the successful changes can outcompete the other less successful ones before the units either (1) die out or (2) mutate so much that whatever made them "fit" to the current environment gets lost (Dawkins, 1982). Figure 1a sketches this idea.

Learning to learn as accumulating fitness

In extending this to the evolution of ideas, the analogy has been made, including famously by Dawkins (1976) in coining the term "meme" as a conceptual-analogue of the basic unit of selection.[2] However, it is less obvious that the analogy is widely accepted in

[2] This launched the field of memetics, studying how ideas evolve. However, as far as we can tell, the field has generally focused at the group level, on cultural transmission as one form of selective retention

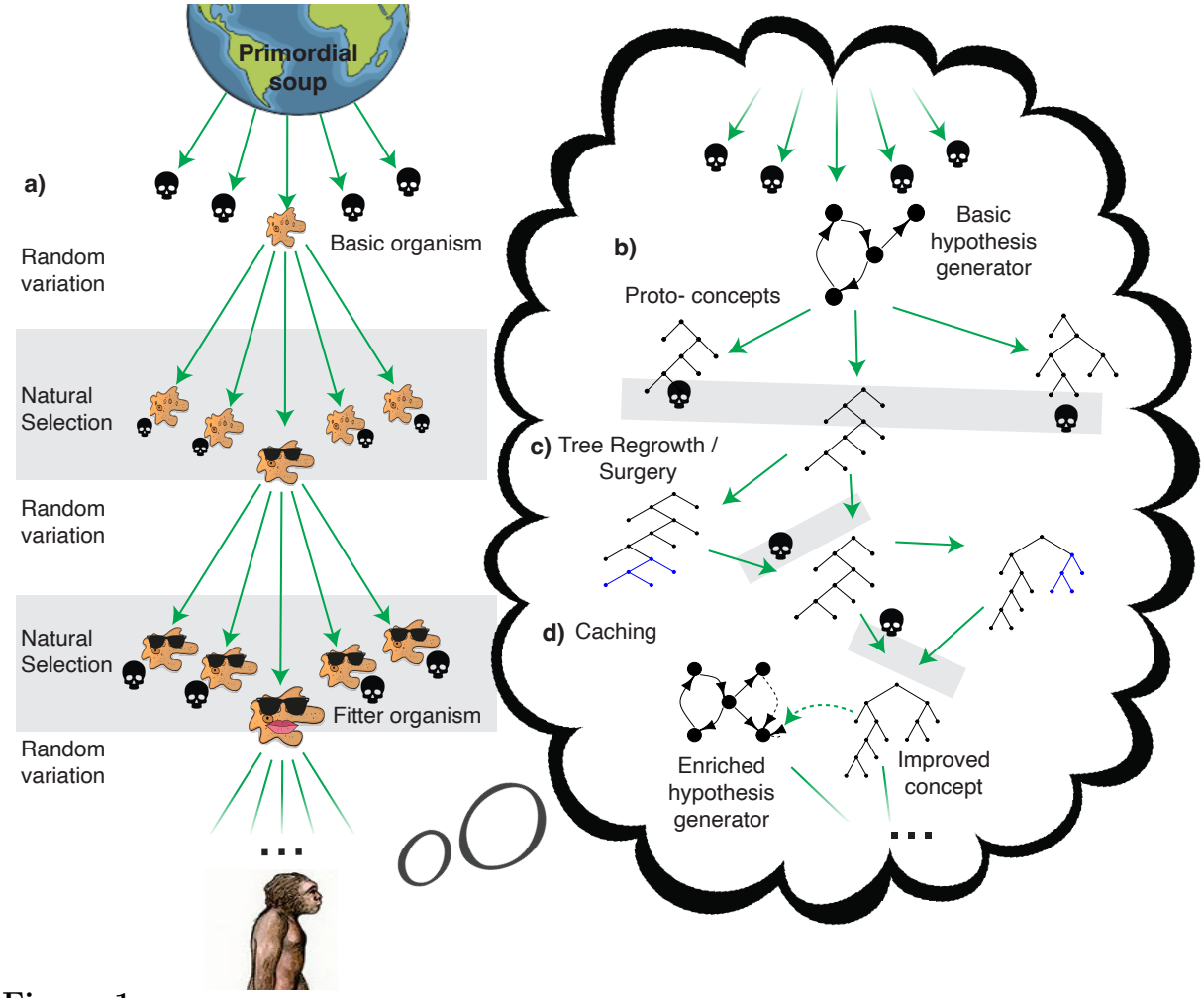

Figure 1

a) Caricature of biological evolution as a long local search via random mutation (green arrows) and selection (gray boxes) starting from a basic replicator, skulls indicate unsuccessful variations. b) A minimal universal concept generator and three possible proto-concept products. c) Tree regrowth/surgery within established concept as a mechanisms of incremental mutation See Figure 3 for worked example. d) Bootstrapping through selective caching of promising concepts for reuse in future concepts. See Figure 4 for worked example.

(Aunger, 2000), and has had thus far had less to say about the mechanisms of variation within the minds of individual cognizers.

mainstream psychology. Neuronal pruning in early development has been characterized as conforming to principles of "neural Darwinism" in which connections compete and better combinations survive (Changeux, 1997; Edelman, 1993; Stanley & Miikkulainen, 2002). But when it comes to individual higher-level cognition, we habitually think of minds as doing something far cleverer, with more goal-directedness and foresight, than the random walk of evolution. Indeed they can do this too. The brain is often described as a kind of prediction machine (Agrawal, Gans, & Goldfarb, 2018; Clark, 2015), inductively forming models that supercharge learning and performance in familiar contexts (Kemp, Goodman, & Tenenbaum, 2010). Much of effort in machine learning in recent decades has been spent in developing systems that better synthesize humans' domain flexibility and data efficiency, by instilling kinds of inductive biases, or established "good tricks" found in human cognition (Lake, Ullman, Tenenbaum, & Gershman, 2017).

On these lines, it seems clear that in any familiar domain, earlier evolution or earlier learning will have fine-tuned the parameters of the search function (both cognitive and environmental), so that possibilities will come to mind that are, at least in the cognizers's Environment of Evolutionary Adaptedness (EEA; Bowlby, 1969; Tooby & Cosmides, 1990), better than strictly random ones. For instance, a child constructing an understanding of the physical world may take search steps that are somewhat stochastic, but that are also biased by a pre-existing scaffolding built by natural selection (Spelke & Kinzler, 2007). A child might be predisposed to act on their environment in ways that are generally effective at revealing its specific causal structure (Bramley, Jones, Gureckis, & Ruggeri, 2022) and latent physical properties (Bramley & Ruggeri, 2022), provided that these predispositions reflect fairly stable properties of their EEA. In general, the cognitive science literature has made a great deal of the idea that we can and do *learn to learn* (Kemp et al., 2010; Lake et al., 2017). That is, we accumulate inductive biases and intuitive theories that get us going quickly in familiar domains (Gerstenberg & Tenenbaum, 2017). However, the success of this strategy depends on our continuing to live in, or close to, the EEA that the approaches

were shaped by. The no free lunch theorem captures the fact that, were these environmental properties and relationships to reverse tomorrow, our inductive biases would only lead us away from what we need (Wolpert & Macready, 1997). For present purposes, we also note that learning to learn necessitates first *discovering* the good ideas or clever tricks we want to reuse when revisiting familiar settings.

Conceptual systems as generative models

In recent decades, the Bayesian, free energy and deep learning traditions have converged on a characterization of the structure of a mature human conceptual system. Roughly minds are thought to encode a hierarchical causal generative model of their environment (Griffiths & Tenenbaum, 2009; Hohwy, 2013; Kemp et al., 2010; Lucas & Griffiths, 2010; Tenenbaum et al., 2006). One can think of this as a cascading constellation of interconnected concepts from very general or abstract ones, such as "cause" and "belief" at the top, to highly domain specific ones like "carburetor", "voter" or "dax". To be of practical utility, the whole collection needs to act as model of the world: compressed, yet sufficiently causally accurate, predictive, and flexible to enable prediction and planning (Conant & Ross Ashby, 1970; Hohwy, 2013). It turns out that a generative hierarchy of causally structured representations, from our most general and universal beliefs to the most specific concepts, fits this bill better than any other structure we know of.[3]

This kind of characterization has helped demystify the things minds do very well, such as perform few-shot inferences (Gerstenberg & Tenenbaum, 2017; Griffiths & Tenenbaum, 2009; Kemp et al., 2010; Zhao, Lucas, & Bramley, 2022), make the kinds of uncertainty-sensitive inferences that early AI struggled with (Clark, 2015; Oaksford & Chater, 2007), and learn proactively by using subjective uncertainty to guide what (inner) hypotheses we want to consider or investigate next, what action to take to resolve where we are and what is happening and so on (Bramley et al., 2015; Gureckis & Markant, 2012;

[3] There are various perspectives on why this is the case (e.g. Badcock, Friston, & Ramstead, 2019) but these are outside scope of this paper.

Nelson, 2005). Most relevant for the current discussion, the generative model framework also seems to capture an important sense in which the mind seems set up to produce stochastic variation and novelty of the sort that could allow for evolutionary mechanisms. We can use a generative world model in a top down way, to sample conditional possibilities for use in inferences (Chater et al., 2020), but also to explore possibilities unconstrained by evidence as we do when we dream or hallucinate (Fletcher & Frith, 2009). Having a generative model seems key to our ability to reconstruct past episodes from a compressed memory trace (Hemmer & Steyvers, 2009), and to play out the counterfactuals that guide responsibility judgments (Quillien & Lucas, 2023) and hypotheticals that guide planning (Guez, Silver, & Dayan, 2012). What is less clear is how the hierarchical constraints, (in)dependencies and structure of this inner state space can come about.

In the next sections we unpack how the notion of learning as program induction, and particularly algorithms developed to perform it, provide traction on how conceptual novelty and complexity arises within a mind.

Mental model building as program induction

We think that recent framings of concept learning as program induction provide a helpful way to think about and formalize the problem of how conceptual structure arises from initially simple mechanisms (e.g. Bramley, Schulz, Xu, & Tenenbaum, 2018; Buchanan, Tenenbaum, & Sobel, 2010; Chater et al., 2020; Dehaene, Al Roumi, Lakretz, Planton, & Sablé-Meyer, 2022; Goodman et al., 2008; Lake, Salakhutdinov, & Tenenbaum, 2015; Piantadosi, Tenenbaum, & Goodman, 2016; Rule, Schulz, Piantadosi, & Tenenbaum, 2018; Ullman et al., 2012). Building on ideas from computer science ideas including inductive programming and program synthesis, program induction methods attempt to automatically generate computer programs to perform tasks, solve problems, or fit data (Church, 1963).

Consider the first time you interact with a new class of objects, and need to learn their properties. Based on stimuli from Zhao et al. (2023), Figure 2a illustrates such a

situation with "magic eggs" that can change the length (number of segments) of "sticks" they touch. Your task is to discover the rule that governs the magic eggs' power on the sticks. This rule can be seen as a simple 'program' that takes as input the properties of an egg and stick and returns a new stick length as output. This simple toy example illustrates an important challenge of learning in most novel situations: The space of possible hypotheses is infinite. This is because one could potentially write an infinite number of programs governing how the magic eggs work, for instance all the following programs are compatible with the observation in Figure 2a:

- the resulting stick length is always one (in pseudo-code: $\text{sticks}' \leftarrow 1$),
- the resulting stick length is randomly generated ($\text{sticks}' \leftarrow \text{rand}(0, 10)$),
- $\text{sticks}' \leftarrow \text{sticks} - \text{spots}$,
- $\text{sticks}' \leftarrow \text{sticks} \times \text{stripes} - \text{spots}$,
- etc

The challenge of program induction is to solve this problem in a tractable way, despite the infinity of potential theories that could explain the data. To do this, program induction works through first defining a kind of meta-program that generates new candidate programs, and a mechanism for searching over these generated programs for those that perform better under a criterion of interest (Summers, 1977).

Under a program induction account of concept learning, a cognizer begins with some set of primitive operations and some mechanism that combines these operations recursively and stochastically until a termination condition is reached. This basic mechanism lays the groundwork, or sets up the space and process that allows for the creation of all sort of "programs" or concepts.[4]

---

[4] We note at this point that by equating concepts with mental programs we are adopting a conceptual role semantics. That is, we are assuming that the meaning of individual concepts is determined by how they connect to and interact with all the other concepts in the mind of the individual. Thinking of programs as concepts therefore commits us to an internalist view of meaning, such that correspondence between someone's conceptual system and the external world is contingent. This means that our account is not

Crucially, anything that can be expressed in the language defined by the primitives, can be generated by the recursive application of such a construction mechanism. This means that there is a basic sense in which everything a program induction mechanism is capable of producing is baked in (Perfors, 2012). Fortunately, even extremely simple computational systems can be highly expressive, such that many can be used to represent and execute any program (Turing et al., 1936). This is because computation unlocks the ability to exploit systematicity and compositionality (Fodor & Pylyshyn, 1988), the same principles that allow the speaker of any natural language to use that language to say more or less anything they like, even if it has never been said before (Chomsky, 1959).[5] Just as biological forms we see today were all merely *potential* arrangements of physical matter long before they became actual products of evolution, all that needs to be built in is a mechanism that has the *potential* to arrive at these programs. Indeed, a variable-free combinatory logic made up of just two terms is Turing complete (Schönfinkel, 1924). Piantadosi (2021) notes that the operations described by these terms are also straightforward to instantiate neurally, making some small set of primitive combinatory logic operations an adequate basis in principle for the emergence of a conceptual system

---

meant to shed light on the thorny philosophical issue of how concepts relate to their external-world referents (Fodor, 1978; Kripke, 1980; Putnam, 1975). For example, our account cannot capture any ways in which concepts have shared public meanings, necessary or accidental properties, nor how individuals might rely on the external world for conceptual details rather than representing them (Rozenblit & Keil, 2002).

[5] Fodor famously argued that on a "language of thought" view, everything the mind is capable of conceiving—from carburetors to string theory—has to be built in from the start (Fodor, 1975). At the time, this was seen as absurd consequence of symbolic accounts of the mind and led to a move away from talking about minds in terms of an inner language of thought. We think the advent of probabilistic program induction techniques makes this claim much less implausible. Baked-in universality is a feature of many simple (programming) languages. By universality, we mean a capacity for generating or representing anything that can be represented using any known system or language; on this account, Turing-completeness can be taken to imply universality and can be achieved with extremely simple generative grammar compared to the complexity of other products of biological evolution.

that includes carburetors and string theory among its potential products.

To formalize program induction, it is common to use Probabilistic Context Free Grammars (PCFGs, Ginsburg, 1966) from linguistics. A PCFG is a set of rules, primitives, and production probabilities defining a language (natural or artificial, e.g., programming) such that, if applied iteratively and recursively, the rules can result in any grammatical statement in that language, and assign that statement a probability. Figure 2a gives an illustrative example of a PCFG imbued with a small set of primitives that allow it to produce concepts about a causal rule that governs an interaction between two abstract objects (Zhao et al., 2023): An agent—a "magic egg" with two quantitative features (spots and stripes)—and a recipient—a "stick" with a single feature, a length, in segments (Figure 2b).

Figure 2c illustrates a sequence of rule applications, here resulting in a concept that is (coincidentally) consistent with the examples in Figure 2d: "The interaction causes the stick's segments to be multiplied by egg's stripes before segments equal to the egg's spots are subtracted from the stick". All other things being equal, a PCFG favors concepts composed with fewer rule applications, implying an inductive bias favoring simpler concepts over more complicated ones.

Monkeys and typewriters

PCFGs have been used in cognitive science in part as a model class for defining priors with support over an infinite space of possibilities (Goodman et al., 2008; Piantadosi et al., 2016; Rule et al., 2018). That is, we can think of a PCFG over a Turing-complete language as expressive enough to ground a computational-level characterization of the grand non-parametric inference problem cognizers face in inducing a good model of the world (Marr, 1982). However, infinite hypothesis spaces come with implications for engineering mechanisms that approximate inference, since it is impossible to consider more than a tiny fraction of the space. A naive approach is to generate a large set of theories from a prior defined by a PCFG and weight these with a likelihood (i.e. fitness) function,

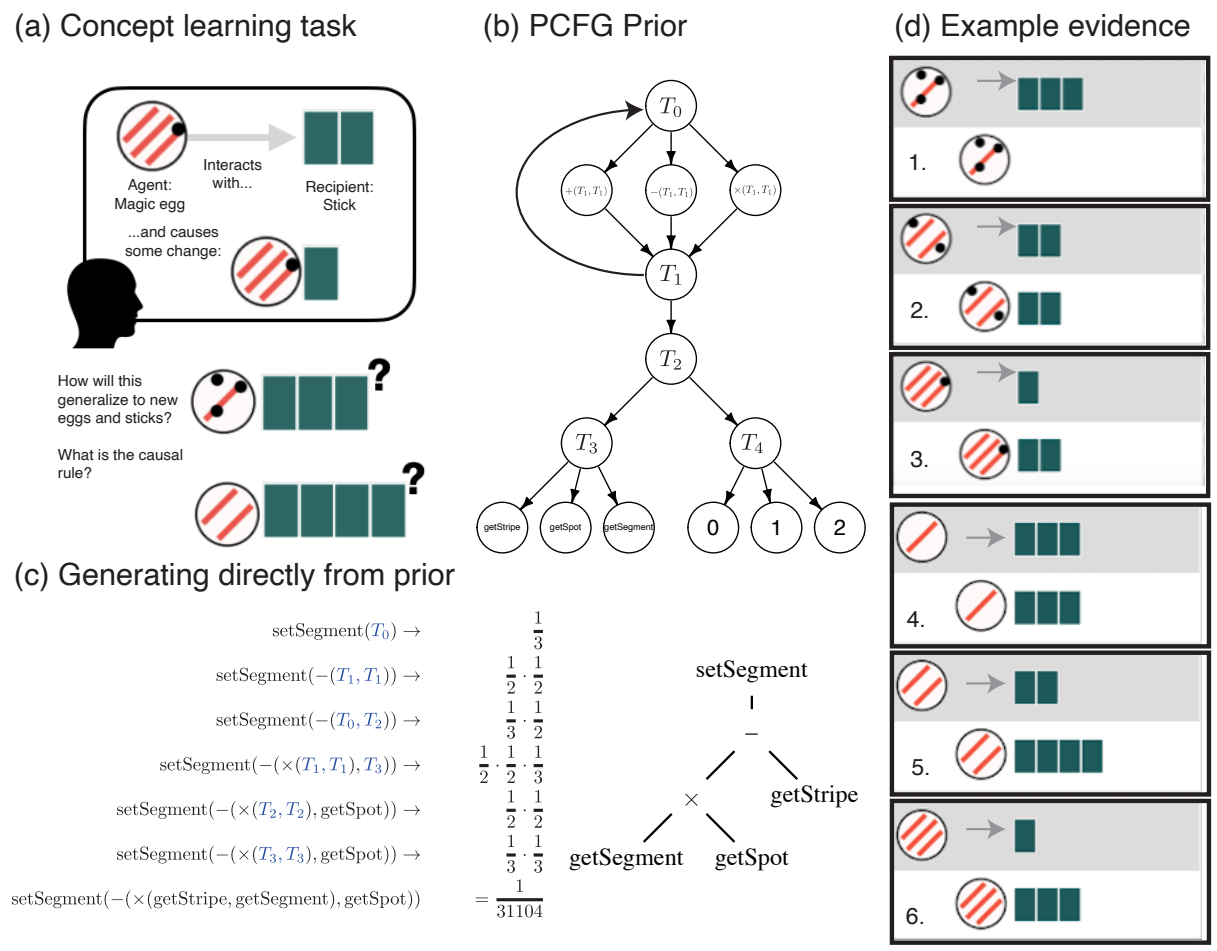

Figure 2

a) A causal concept induction scenario b) Illustrative PCFG covering a space of possible causal rules. Starting from T_0 one follows arrows at random recursively replacing any non-terminal placeholders ($T_0 \ldots T_4$) at the arrow's source with the content at the arrow's target. c) Example of generating a concept directly from the prior (consistent with d only through blind luck). Each line replaces the non-terminal T_n with one of its possible productions with the fractions tracking the probability of each. d) Six data points.

so that the weighted sample acts as an approximation to the posterior over concepts given evidence. Unfortunately, this runs straight into the "monkeys and typewriters" problem we described earlier in relation to biological evolution (Borges, 1941/1998). All concepts or programs can in principle be generated directly from such a universal prior, but in practice

the probability of this happening for complex concepts is negligible. As an illustration of this, in the toy scenario illustrated in Figure 2a, assuming a uniform probability of taking each branching path, outputting the concept that in fact governs the causal effect responsible for the six learning examples, has a probability of $\frac{1}{31{,}104}$, meaning one would need to generate many thousands of prior samples to have a good chance of discovering it. Worse, even for this toy scenario this PCFG is too simple. When participants were tested on this example by Zhao et al. (2023), many reported rules including "and"s, "or"s and conditional "if/then" statements. Enriching a PCFG such that it can also generate these guesses, e.g. including Booleans primitives to allow for anything in first order logic, the chance of generating even this simple ground truth becomes astronomically small. Fortunately, this computational challenge has driven the program induction research community to develop other tricks for efficiently exploring the state space of models. We highlight two such approaches and explain how both rely on incremental blind mutation and selective retention. We argue this helps explain how conceptual complexity can grow and evolve gradually within a cognizer's mind.

Algorithms for program induction

Stochastic local search

Markov Chain Monte Carlo (MCMC) methods are a widely used statistical approach for the systematic exploration of intractable state spaces (Brooks, 1998). They produce chains of samples, or hypotheses, that have equilibria approximating a distribution of interest, such as a posterior that cannot be computed directly. One reason MCMC methods are an interesting approach for our purposes is they describe a principled stochastic local search over possibilities typically for use within spaces where direct optimization is impossible (Suchow et al., 2017). Moreover, they achieve this via the same ingredients we have argued are central to biological evolution: sequences of simple random variations combined with forces of selection. For example, in the Metropolis-Hastings MCMC algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), blind

variation comes from the proposal function. This can be a range of modifications of the current hypothesis but typically selected to be something easy to generate, and neither too large—as to depart completely from the current hypothesis—or too small—as to make the chain practically static. Selection is then achieved through an acceptance function, that stochastically accepts or rejects a modified hypothesis depending on its fitness relative to the current hypothesis. When a relatively better hypothesis is encountered, such a mechanism tends to accept and then retain it for longer because most subsequently proposed mutations will be rejected. When a bad hypothesis is adopted, the process will tend to depart from it quickly because most proposed mutations will be accepted. Provided the proposal function makes all parts of the space accessible, and the acceptance function is based on both prior (i.e. complexity) of the hypothesis and its likelihood (i.e. fit to purpose), then the resulting chain of hypotheses will eventually visit every possibility with a frequency proportional to its posterior probability. If one instead always accepted the better hypothesis, the sequence of hypotheses would tend to improve quickly but then get stuck in a local optimum where all local proposals are worse. If one instead randomly chose whether to accept each hypothesis, then the algorithm would simply walk from one hypothesis to another without favoring those that fit better. MCMC thus occupies a sweet spot in the space of stochastic search algorithms between two maladaptive extremes: A level of randomness that allows a cognizer to explore concept space without ever getting completely stuck, but also to favor the parts of the space with higher fitness.

Many recent Bayesian models have used MCMC to make human-level and human-like inferences in a variety of concept learning settings (e.g. Goodman et al., 2008; Lake et al., 2015; Piantadosi et al., 2016; Thaker, Tenenbaum, & Gershman, 2017). In these models—and in most applications of MCMC in the cognitive sciences—MCMC is used as a tool to approximate true or optimal posterior distributions, typically in service of a computational-level theory, and typically compared with group-level behavioral data; it is not offered as a specific process-level claim about how individual participants navigate the

inferences in question, as we do here. However, the idea that MCMC-like mechanisms describe the workings of the mind itself has also been suggested (Abbott, Austerweil, & Griffiths, 2015; Bramley, Dayan, Griffiths, & Lagnado, 2017; Castillo, León-Villagrá, Chater, & Sanborn, 2023; Dasgupta, Schulz, & Gershman, 2017; Lieder, Griffiths, M. Huys, & Goodman, 2018; Sanborn, Griffiths, & Navarro, 2010). For instance, Dasgupta et al. (2017) use MCMC to explain a variety of response biases in conditional probability judgments such as "what is the probability that an image containing a table also contains a [chair/computer/curtain]". Lieder et al. (2018) similarly propose cognizers adjust quantitative estimates (e.g., about the arrival time of a bus) away from an arbitrary initial seed or anchor via MCMC. Ullman, Stuhlmüller, Goodman, and Tenenbaum (2018) propose people adjust an initial summary-statistic derived estimate of physical parameters through a short MCMC chain over parametrisations within an intuitive theory of physics. Castillo et al. (2023) explain biases in random number generation as stemming from limited local search with momentum. Davis and Rehder (2020) explains biases in causal model based judgments as resulting from limited exploration of the state space of those causal models anchored on their most characteristic canonical states. Bramley et al. (2017) explain sequential dependence in people's causal structure judgments across learning instances as resulting from people randomly mutating their causal models through short MCMC-like chains to search for alterations that accommodate new evidence. Fränken, Theodoropoulos, and Bramley (2022) applies the idea to a Boolean concept learning task somewhat similar to Figure 2a, contrasting process models based on several stochastic search proposals to explain autocorrelations in guesses that people make.

Focusing on how MCMC can be applied to compositional symbolic spaces such as those defined by a PCFG prior, one well established proposal distribution is "tree-regrowth" (Goodman et al., 2008), which represents a concept as a tree-like structure where each node is a function taking arguments from the leaves below it and passing the result to the branch above (e.g. Figure 2c). Randomly deleting a node and everything

beneath it, and replacing it with the non-terminal that produced it, leaves a concept with a "gap" that needs to be filled. The gap is filled by using the PCFG to eliminate this and any newly created nonterminals, and arrive at a new proposal (see Figure 3). We imagine the learner starts by entertaining the possibility that the stick grows through multiplication by the egg's spots followed by addition of its stripes. They randomly select the "×" node of the concept tree, delete it, and regrow a new subtree until termination using their PCFG. Fränken et al. (2022) use short MCMC chains of these kinds of tree regrowth proposals as a candidate process model explaining how learners go from one guess to another as they reason about a geometric Boolean concept. However, they find human patterns of sequential dependence between participants' guesses to be both different and stronger than those predicted by tree regrowth. For example, participants often made changes to "upstream" nodes of a concept without also making downstream changes. Indeed, the first hypothetical mutation results in a candidate concept—that the stick will grow or shrink to be of length 3 minus the spots, plus the stripes—that is almost completely different from the previous one, as well as being a poor characterization of the data in Figure 3b, so is not selected.

To travel from their starting hypothesis to the correct hypothesis, the "+" at the root of the tree needs to change into a "-" without disturbing the rest of the concept. To allow for this kind conceptual move, Fränken et al. (2022) propose a more conservative proposal distribution they call "Tree Surgery". This involves a handcrafted localist proposal mechanism that randomly replaces, adds, removes, or splices in one new node at a time. They found that short MCMC chains based on this proposal distribution did a better job of capturing participants' conceptual revisions than Tree Regrowth and several heuristic search mechanisms. Illustrating this, we show a surgical mutation in Figure 3c which improves the quality of the hypothesis and so is accepted.

Across these examples, the key signature of MCMC-based process models that distinguishes them from the normative models they approximate, is their sequential

dependence, producing patterns of auto-correlation and anchoring in a single cognizer's ideas over time. For Dasgupta et al. (2017) and Lieder et al. (2018) this explains participants' dependence on context and prompts, while for Bramley et al. (2017) and Fränken et al. (2022) it is on the learners' own earlier-reported belief. This is because a short chain of MCMC steps will tend to remain close to where it began. Autocorrelation order-effects are so ubiquitous in psychology experiments, that we take pains to average them out of data collection through counterbalancing and statistical averaging. Conceptually, we should expect this autocorrelation to increase as we scale this idea and these results up from toy experimental scenarios to the larger problem of adjusting one's global beliefs about the world. That is, if we imagine a cognizer's entire ontology, or world view, as a very large connected, tree-like graph, then the entire model is a point within an infinite latent possibility space of possible worldviews and the idea they could generate a completely independent alternative is tantamount to restructuring their entire cortex. In addition to being practically infeasible, this leads us back to the monkeys and typewriters issue, namely that a complex random proposal has a negligible chance of having good fitness. Consequently, the idea that mature cognizers are limited to much smaller, more local and incremental modifications seems inevitable. Elsewhere we have argued that this is related to the "antifoundationalism" captured by the Duhem–Quine thesis and the metaphor of *Neurath's ship* (Bramley et al., 2017; Duhem, 1954; Quine, 1969). The Duhem-Quine thesis posits that no scientific hypothesis stands or falls in isolation because any attempt to test it will depend on auxiliary hypotheses and assumptions. In a similar way, it seems like almost any theory revision in the mind of an adult is bound to be conditional on the wider network of beliefs in which the revised element is embedded, e.g. that their senses are working as expected (Gershman, 2019a). The antifoundationalist consequence of this view is captured in the Neurath's ship idea: Theory change is rather like patching a ship while at sea, with each change depending always on the current global hypotheses (the rest of the ship's hull) for support, without the privilege of foundation on

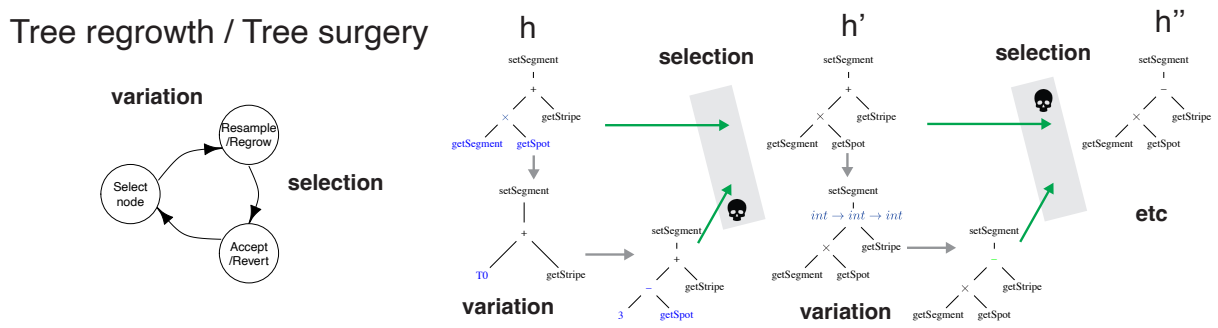

Figure 3

Illustration of Tree Regrowth MCMC and Tree Surgery MCMC. Grey arrows represent the proposal of new trees by blind variation; green arrows represent selection. Learner starts with hypothesis h, proposes a modification by randomly selecting and regrowing the bottom left branch (an example of Tree Regrowth). This is rejected, making h′ the same as h. Learner then surgically replaces "+" with "-" (Tree Surgery). This improves fitness, so is accepted, arriving at the target concept.

which to disassemble and rebuild the entire model (as one could in a shipyard or dry dock).

Supercharging local search

We have so far suggested that MCMC-like stochastic search over a PCFG-defined concept space gives some traction on the puzzle of how minds might incrementally grow and adapt their inner theories and world models. However, the PCFG seems too simple a mechanism to capture gamut of human concept change. The analogy of concept change as MCMC over a PCFG is limited by classical MCMC mechanisms' inability to compress data, as well as the PCFG model class's fundamental lack of modularity.

(Lack of) Compression. An MCMC-PCFG chain describes a walk around concept space that is designed to have a specific stationary distribution. Treated as a model of inference, it does not predict that cognizers will spontaneously progress from simpler to more complex concepts, except by accumulation and evaluation of more data over time. An MCMC chain, by design, has no memory for what other ideas it has landed

on in the past beyond whatever hypothesis is currently in place.[6] As more evidence arrives this will affect the search steps such that ability to explain the data increasingly outweighs the prior preference for simplicity (Howson & Urbach, 2006), licensing increasingly complex models be accepted by mutations. However, this requires storing and evaluating ever more data. This is clearly unworkable as an account of life-long learning, where a key role of representations is to absorb data so it can be forgotten.

(Lack of) Modularity. PCFGs allow for a limited form of learning. One can learn production probabilities from data, e.g., by introducing a Dirichlet prior over the possible expansions of a particular symbol in the grammar. For instance, if in our previous example "+" has featured in more successful concepts in the past than "×", we can infer that the probability of producing "+" should be higher than "×" when generating or mutating concepts henceforth (Figure 2). However, the fact that production probabilities in PCFG are independent of context mean the scheme cannot learn how to better *combine* concepts. If learner is fortunate enough to land on a powerful concept or "good trick"—i.e. some clever composition of their primitives—under a PCFG-MCMC scheme, the combination is only retained as long as it continues to feature in their hypothesis, i.e. survives subsequent random mutations. Under a Tree Regrowth proposal function, for example, half of a learner's concept tree is regrown on average with every mutation, meaning that nontrivial discoveries will tend to be obliterated by subsequent mutation. This is illustrated by the example in Figure 3a. The learner starts their search syntactically close to the target concept. That is, they have already generated a key piece of the target concept in identifying the multiplicative relationship between the stripes and segments. However, because the discrepancy between their hypothesis and the target is near the root (where they have a "+" instead of a "-"), they cannot modify their hypothesis and arrive at the target concept without also "rediscovering" the lower branches by generating them

[6] Although, see (Quiroz, Kohn, Villani, & Tran, 2018; Welling & Teh, 2011) for examples of modern MCMC algorithms that do have some capacity to compress data.

anew from the prior. Tree Surgery mitigates this problem being much more conservative in its mutations, but does not eliminate it.

This illustrates a fundamental limitation of a PCFG based model of inference. It implies the learner is forever stuck drawing on their original set of primitives. If the cognizer's experience suggest the same complex concept applies in several situations, that concept must be rediscovered each time. But re-use of useful fragments is fundamental to the evolution of both biological forms and thought. Biological evolution often innovates by making variations on the same underlying motif: we have 10 fingers, but did not have to independently evolve each of them from scratch 10 different times (Carroll, 2005). The ability to repurpose tricks wholesale seems to be quite important for human cognition. Indeed, to a first approximation, this is what it means to reason by analogy (Gentner, 1983; Holyoak & Thagard, 1996). Therefore, a key part of characterizing a conceptual evolution seems to be explaining how we can reuse discovered structure. This is where the idea of bootstrapping a library of concepts and the formalism of adaptor grammars comes into play.

Bootstrapping

Bootstrapping—the paradoxical notion of "lifting yourself up by your bootstraps"—classically refers to transformative learning where the "the endpoint of the process transcends in some qualitative way the starting point" (Carey, 2004, p59). In statistical practice, it has acquired a more specific meaning, referring to inference techniques where a model's output is fed back into the model as part of a training or inference loop. By using a generative model capable of storing and reusing its past ideas, we can reify the statistical principle of bootstrapping and apply it to concepts. This requires using a representational framework that overcomes the limitations of standard PCFGs. One recent approach that achieves this is the adaptor grammar formalism (Johnson, Griffiths, & Goldwater, 2006; P. Liang, Jordan, & Klein, 2010; Zhao et al., 2023).

A helpful analogy for thinking about how an adaptor grammar works is learning to

program (Rule et al., 2020). A novice coder might only know a few very basic functions and initially find that coding anything interesting becomes repetitive, tedious and error prone. But as they become a more proficient coder, they will learn to store useful chunks of code as reusable functions that take, and return particular types of variables (`and(x,y)` combines two Booleans into a third Boolean, `greater(x,y)` combines two integers into a Boolean etc). As a basic example, if you frequently need to compute averages it would be more efficient to define a "mean" function where e.g. `mean(x) = sum(x)/length(x)` and call this time each time a mean is needed rather than writing out the operation over and over again. In general, the history of programming language development has this character. Modern languages like python and R provide large libraries of powerful functions and abstractions that compile down to lower level programming languages, bottoming out in a binary machine code.

In order to enable modularity in a program induction framework, we need a formalization that maintains some separation between structure and functional form. Combinatory logic (Schönfinkel, 1924) provides this by using a system of terms and types to to combine functions. Router terms govern how values flow through the computational tree structure. This removes the need to keep track of variables and so makes it far easier to glue programs and subprograms together. In combinatory logic, each variable comes with its type—e.g., *int*, *obj*, *bool*—and for each functional term we write its types in the order of arguments and then outputs, such that `greater(x,y)` has "type signature" $int \rightarrow int \rightarrow bool$. In our illustration, we use several primitive functions and four router terms **B, C, S, I** for expressive convenience, but they can be re-described as combinations of just two base terms **S, K** (Schönfinkel, 1924), see Figure 4a and Zhao et al. (2023) for details. To parse the program depicted in Figure 4a the egg object is passed in first and routed right by **B** (the first router at the root of the tree), then to both sides by **S** ultimately being fed to the getSpot and getStripe primitive functions at the right hand leaves. Meanwhile, the stick is passed in second, so routed left by the **C** at the root, right

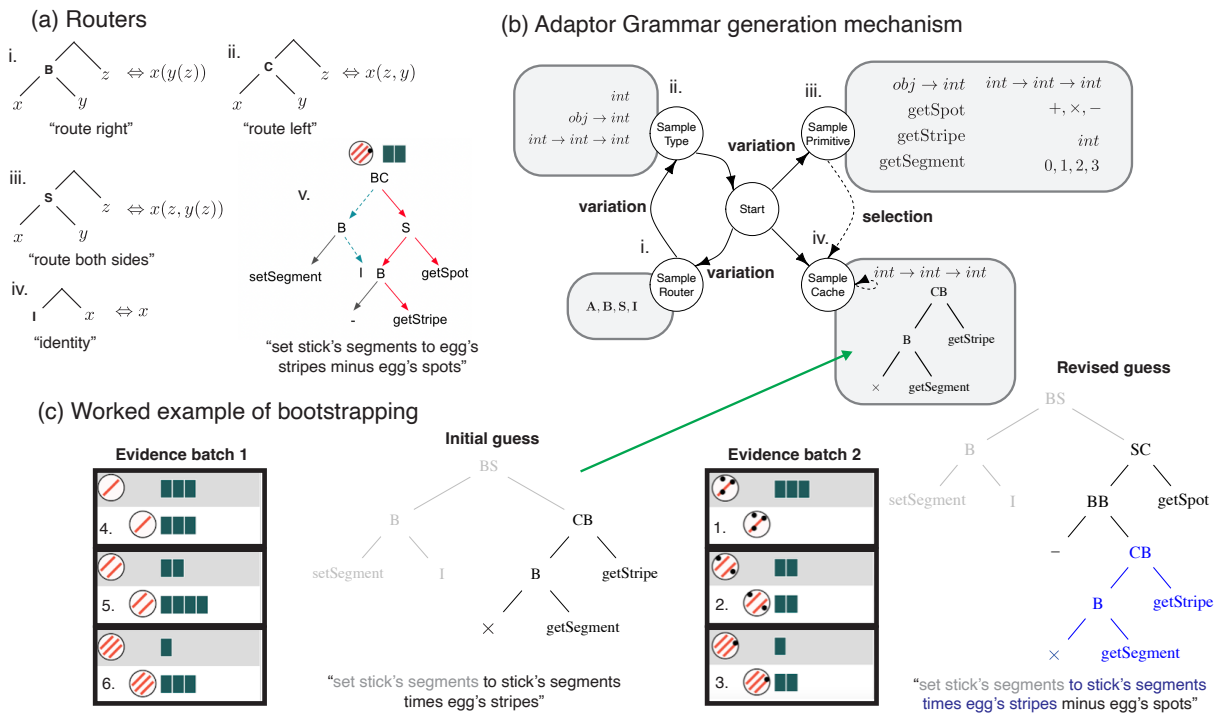

Figure 4

(a) Expressing a modular variable-free concept using routers: i–iv The four routers. v. Worked example where red solid arrows indicate where the egg is passed and teal dashed arrows indicate where the stick is passed. (b) Adaptor grammar: Learner recursively samples i. Routers (creating new branches and determining inner structure of concept) and ii. type signatures for each leaf governing the permitted routed arguments, then base terms either primitive (iii.) or cached (iv.), terminating when no branches remain. (c) Illustration of how complete programs can be selectively added to the cache and subsequently be selected as base terms.

by the **B** below and then to the leftmost leaf by **I**. The complete expression thus sets the stick's segments to a length determined by the egg's stripes minus its spots. With this framework in place, modular mutation and reuse becomes straightforward. For example, the getSpot leaf can be replaced by any term that reads an object as input and returns an integer as output, thanks to routers and type-constraints. A Tree Surgery type search scheme is thus perfectly feasible here as well.

With the notion of routers and types in hand, we can now think about how to use AG to generate and cache concepts. The process begins with the target definition type (Figure 4b): What kind of input will the program get and what kind of answer does the program being generated need to return? This is represented by the routers at the root of the tree. In our example we are searching for a program that takes the two objects (egg and stick) as input, and returns a stick (with potentially a different number of segments) as output. With $\texttt{setSegment}_{obj \to int \to obj}$ on the left hand side of the tree, the right hand side sub-tree provides an integer that `setSegment` uses to alter the length of the stick in ways that hopefully also explain the evidence. The process proceeds by recursively growing out the leaves of the tree either by sampling a router (creating more leaves), or a type-appropriate base term, until all leaves are filled with base terms. When complete, the expression will have routers all internal nodes and base terms at all leaves. Crucially, whenever AG finishes growing a concept, it can cache it under its type signature, and add it to the library of base terms that it will draw on in future (Figure 4c). The next time the learner goes to make an inference, their prior ideas are now included in their library of primitives they can draw on, making it straightforward to reuse previous conclusions wholesale in a new expression without re-inventing them each time. Whenever a concept is cached that is already in the learner's library, we can either think of the library as containing several copies of that term, or more succinctly as increasing the selection weight on that term. Free parameters can also govern the degree of reuse of cached rather than primitive base terms, as well as governing the balance between growing branches with new routers or terminating them with base terms, similarly to the weights in a PCFG. To prevent such a library from growing unboundedly, one can implement storage limits such that cached terms that do not prove useful eventually fall out of the library and are forgotten. While the original use case of AG was to learn a model of the sharing and reuse of sub-structures in language, here we repurpose AG's cache-and-reuse mechanism to formalize the idea of bootstrap learning.

Returning to our running example, consider the evidence in Figure 2d. Even though this is a toy problem, it is challenging to find a hypotheses that fits all of these examples. However, if you focus first on examples 4–6 where there are no spots, it is relatively easier to come up with an partial explanation: E.g. perhaps the stick's segments get multiplied by the egg's stripes. If one infers this and caches it, then proceeds to reason about all six examples, you can now reuse the cached insight as a new $int \to int \to int$ primitive, and so more easily discover the intended concept which requires this fragment to be nested within the subtraction of the egg's spots (Figure 4d).

Zhao et al. (2023) ran a series of experiments in which participants had to make judgments about causal concepts like these based on evidence like that shown here presented in different sequences. In one condition, participant were shown examples 1–3 (where the eggs have both stripes and spots) and asked to make a guess, before being shown items 4–6 and being asked to make a revised guess explaining all the evidence. In this condition, generalization accuracy was close to chance and no participant described exactly the intended concept. Participants rather tended to describe complex and poorly fitting concepts. For example, several participants proposed that perhaps: "dots remove segments, stripes add segments, excepting when there is just one stripe, in which case nothing changes" (explaining 4 of the 6 examples). Zhao et al show that this pattern is to be expected under the model we sketch above, if we assume that learners cached whatever they had inferred from reasoning about just the first three items, and attempted to generate a concept that could capture all six items. A separate group of participants were shown items 4–6 first before being shown items 1–3. Strikingly almost half the participants were then able to go on to guess the intended concept correctly and generalization accuracy (guesses about what would happen under different egg–stick combinations) doubled from 22% to 45%.

Crucially, an AG learner carries a growing library of interconnected concepts rather than only carrying forward just a singular current belief, as in the PCFG-MCMC scheme.

This makes their representation richer and so better able to absorb and retain insights from historic data. The concepts in the library naturally have hierarchical relationships with one another, with older concepts featuring deeper in the stack, as subcomponents of more recent concepts. This naturally provides a route through which hierarchical structure can grow within a mind. Of course our mature conceptual systems are not just *chronologically* hierarchical. They are hierarchical in ways that exploit the potential for compression via abstractions. That is more general features that are true of lots of concepts, appear deeper in the hierarchy. Fortunately, another strength of the combinatory-logic-plus-AG formalism is that it enables a variation and selection mechanism for rearranging this hierarchy. The modular structure of the concepts embedded in an AG library allows them to be "refactored" (P. Liang et al., 2010). Roughly speaking, the leaves of the tree can be shuffled around along with their routers, without altering the meaning of the whole expression but potentially increasing the compression component of concept fitness. This allows yet another stochastic mechanism of incremental search to improve on the happenstance of chronological concept construction (cf. Felsenstein, 1974, for a similar role for sexual selection). In this way, sub-concepts can be collected together and re-cached in different arrangements, unlocking the ability to discover and factor out powerful abstractions (Dechter, Malmaud, Adams, & Tenenbaum, 2013; Ellis et al., 2021; Gershman, 2017; Tian, Ellis, Kryven, & Tenenbaum, 2020).

Summary

Putting MCMC and adaptor grammars notions together and foregrounding their relationship with local blind evolutionary search, we arrive at a sketch of a concept learner with multiple complementary mechanisms for generating hypothetical variation and selectively retaining this variation. MCMC search describes "lateral" variation selected conditional on fit, while bootstrapping captures a recursive variation of modules or chunks selectively retained as reusable building blocks. The accumulation of more powerful, but directly reusable, chunks as primitives, captures how, as we develop, our reach can grow (in

terms of the complexity of the hypotheses we operate on and the changes we consider), even as our capacity for search over potential mutations to our ideas might stay roughly constant. It is important to keep in mind that neither local search nor bootstrapping magically solve the no-free lunch theorem. There is no guarantee that fitter designs will be local to our current beliefs, nor that the concepts one has cached during learning will turn out to be the right building blocks to build more powerful concepts. In fact, if one caches the wrong concepts early during learning, this will only make it harder to arrive at a good trick (Wolpert & Macready, 1997). We have argued in this paper that this is a general feature of learned strategies, intuitive theories and inductive biases: They commit to a certain interpretation of reality that is always in danger of being usurped. The inductive setting thus demands a fundamental flexibility in a cognizer that we have equated with the functional use of localized randomness to serve blind mutative search over symbolic structure of inner models or theories. The path dependence in an individual's beliefs that results from this is evident in our richly developed and deeply ingrained core beliefs and capabilities, packed with tricks for exploiting our environment but also constraining and localizing our new ideas to those that build on or around what we already believe. This is demonstrated on a much smaller scale in our running example from Zhao et al.'s (2023) experiments. Participants rarely generate the target concept in when viewing evidence in the order it appears in Figure 2d, instead getting tangled in complex disjunctive hypotheses that make it harder to get the answer. When experiencing the same evidence in reverse, almost half of participants establish the necessary inner sub-concept and then the intended compound concept. The paper shows that not only these accuracy patterns but also participants' specific idiosyncratic mistaken ideas can be synthesized by the bootstrapped search scheme we sketched here.

Discussion

In this paper we engage with the question of how people develop new theories or models of the world. The problem is puzzling because our concepts seem to live within an

infinite space of symbolic compositional possibilities that can never be fully explored. In line with evolutionary theory, we suggest that sophisticated concepts can be reached via a simple mechanism of random, or "blind", variation and selective retention. Furthermore, accounting for the complexity of our mature world models demands a model that is able to accumulate these variations incrementally, such that many small innovations can ratchet over development as in biological evolution (Felsenstein, 1974). We described two algorithms from the program induction literature that capture elements of how minds could achieve this. Specifically, we highlighted MCMC-like local tree search over models, combined with bootstrapping to "lock in" promising sub-concepts for reuse. Generally, we argued that a consequence of the evolutionary perspective is that, at the cutting edge of cognition, innovation is necessarily: blind (D. T. Campbell, 1960), local (Bramley et al., 2017), and path dependent (Zhao et al., 2023).

We now discuss three key challenges for fleshing out an evolutionary theory of cognition and finally highlight what we see as three key implications of this perspective.

Three challenges

Defining fitness

We have argued that ideas from computer science and statistics (such as MCMC) provide models of how the mind might build theories and concepts incrementally. Under these models, the fitness of a theory is its posterior probability: It is determined by the likelihood that the theory assigns to the data, as well as the theory's prior probability. However, the relationship between the accuracy of one's model and the fitness it conveys need not be one-to-one (Sharot, 2011; Szollosi & Newell, 2020), moreover the criteria the mind uses for selective retention might easily depart from likelihoods and priors. To a first approximation, the fitness of a discovered theory or model is related to how it helps the organism solve the adaptive challenges of their environmental niche. For example, our concepts facilitate mental substitutes for costly physical actions, such as substituting environmental exploration with model-based planning (Ashby, 1952) and the various other

forms of learning-to-learn we describe above. Since we know good regulators should be accurate this would seem to drive selection toward models that are likely in the Bayesian sense of explaining the data. Intuitively, the other component of concept of fitness is the complexity of the concepts that emerge, with more complex theories simply being harder to generate, store or use in downstream tasks (Chater & Vitányi, 2003), lining up approximately with a "prior" preference for simplicity. Importantly, the notion of "prior" complexity also evolves within an AG framework. When the system caches a complex conceptual discovery, it can later use it directly, creating a ratchet effect whereby progressively more sophisticated conceptual constructs become easily available via the same small local mutation steps.

Whither optimality?

A consequence of evolutionary accounts is the implication that the relationship between one's current belief system—both bits that are built in and bits that are learned—and the unknowable ideal generative model or ground truth is always going to be an "unknown unknown" (Knight, 1921). Just as we cannot imagine what the optimal organism would look like, we cannot imagine, claim to already possess or already have a model of, an optimal mind. If we did, we could simply build or adopt it.

The algorithms we highlighted come from probabilistic modeling in machine learning. This seems to cohere with the idea that the brain could be seen as a kind of generic Bayesian sampler (Chater et al., 2020). That is a slightly stronger claim than we are making here. The idea that minds induces novelty via mechanisms of blind variation and selective retention is compatible with but not committed to the selection processes being well calibrated to respect probability theory. That is, minds need not come with an MCMC search algorithm or adaptor grammar "built in" for useful structure and concepts to emerge during development. A better way to think of it might be that these algorithms describe mental programs for novelty generation that are useful "sweet spots" in a larger space of variation–selection mechanisms, making them something that cognition might

discover and subsequently exploit. Along these lines, we might think of the metropolis rule or selective caching, as mental programs discovered by more basic search mechanisms and retained as useful tricks that avoid the "dutch book" inconsistencies that occur when one is insensitive to probabilities (Oaksford & Chater, 2007), and so supercharge a mind's capacity to land on more concepts that compactly reflect the evidence from the environment. This relates to recent proposals about how specific learning algorithms can emerge from more basic forms of learning (cf. Andrychowicz et al., 2016; Dasgupta, Schulz, Tenenbaum, & Gershman, 2020).

Interfacing with subsymbolic processes

Program induction is extremely computationally demanding. Because in its most general form, it implies the learner starts from a universal prior over programs, favoring nothing but simplicity, it can take a very large amount of brute force search to land on good solutions to specific problems (Ullman et al., 2012). This has meant that, in practice, program induction demonstrations of any scale have had to find ways to dramatically narrow the search space. One approach is to use neural networks to compress high dimensional inputs, like images, into a smaller set of abstracted discrete features for program induction to work with (Fränken, Lucas, Bramley, & Piantadosi, 2023; Mao, Gan, Kohli, Tenenbaum, & Wu, 2019). The challenge here is that the two problems are interdependent—the search over programs depends on the feature space while the choice of features depends on their utility in the programs, thereby demanding a joint optimization (cf. Y. Liang, Tenenbaum, Le, et al., 2022). As a related approach, Dreamcoder (Ellis et al., 2021) describes a general purpose architecture that can be trained to master a diverse set of program induction tasks from raw (i.e. pixel level) inputs. The training works by switching back and forth between learning a domain specific language, or generative prior, for the specific task—using algorithms like those we describe in this paper—and training a neural network to approximately invert this generative model so as to propose hypotheses to explain data directly. While data-driven proposals have their own limitations, tending to

"overfit" whatever patterns are in the data (Bramley, Rothe, Tenenbaum, Xu, & Gureckis, 2018; Michalski, 1969), a balance of (symbolic) prior-driven and (sub-symbolic) data-driven computation may well produce good solutions more efficiently than solely searching generatively over programs as in the algorithms we described. In sum, neurosymbolic hybrid approaches blend sub-symbolic and symbolic mechanisms in interesting ways that may yield new insights as to how these processes are blended in cognition. In line with what we have argued throughout the paper, the neural network component plays a *localizing* role, similar to that of the production weights in a PCFG, the starting hypothesis in the MCMC scheme, and the concept library in an AG scheme, biasing the generation of new hypotheses toward concepts that are, in some sense, close to the discoveries made earlier during learning.

Three implications

We finally outline three key implications of this perspective: (1) for making sense of thinking; (2) for supporting learning and teaching; and (3) for understanding our own limitations.

What we are doing when we are "thinking"

One thing the evolutionary perspective helps explain is what people are doing when they are ostensibly "off task". We spend large portions of our time doing things like sleeping, mind wandering, as well as what we might call idea generation, problem solving, or "brainstorming" . Offline learning is a challenge for normative accounts of cognition since it seems to result in haphazard belief change without new evidence. However, from an evolutionary perspective, cognition has plenty of work to do "offline". One way to think about the generation of new ideas is as analogous to prospecting for gold. Since we do not know where the gold is buried, prospecting requires a combination of hard work and luck. It starts with search (panning rivers for dust) but progresses through commitment, e.g. once one starts digging a mine to investigate a possible seam. Since most undiscovered gold is underground, those who strike it rich probably put in the work, and also focused their

efforts selectively, digging deeper where more gold dust could be found at the surface. Analogously, if one is to discover better concepts—new maxima in an infinite and thorny space of possibilities (cf. Ullman et al., 2012)—a lot of energy will need to be spent searching—generating, entertaining, adopting or discarding possibilities—but also through committing to dig below the surface—caching and reusing subconcepts. We have argued our complex concepts are discovered through the combination of stochastic search (as in MCMC) and commitment (as in bootstrapping). We think this gels with the recent suggestion that noise in cognition is better thought of as an essential feature than as a bug (Sanborn et al., 2022), unlocking our capacity for creative thought and growth.

Insights for teaching and curriculum design

The existence of local search and bootstrapping mechanisms in the mind, has a variety of consequences and interactions with applied questions of what makes for a supportive cognitive niche (Clark, 2006; Tooby & DeVore, 1987) and how should we design curricula to help learners reliably "install" complex concepts. In the causal concept induction case we used as a running example, Zhao et al. (2023) demonstrate the importance of the order in which learners attend to examples. It is clear that existing pedagogical principles reflect these ideas to some degree: We know to teach simple concepts at first and built up to more challenging ones. But a formal model of conceptual bootstrapping could help optimize curricula for particularly challenging concepts and help diagnose specific failures in learning and tailor individual corrective curricula in areas such as math and science education (Rafferty, Brunskill, Griffiths, & Shafto, 2016; Rafferty, LaMar, & Griffiths, 2015). The program induction perspective has potential to provide insights for how we actively explore our environment when we don't already have a model of our uncertainty about it.

One opportunity to surpass our individual limitations seems to arise at the population level. If we think of individuals' belief trajectories as particles, cultural evolution can be seen as acting like a kind of fitness-promoting filter (Daw & Courville,

2008). As such we might think of a social dynamics and communication as implementing another quasi-random approximate probabilistic inference scheme, such that individual learners can adopt, repurpose, and modify each others ideas as well as mutating their own (Hawkins et al., 2022; Morgan, Suchow, & Griffiths, 2022; Vélez, Christian, Hardy, Thompson, & Griffiths, 2023). A large enough, diverse enough group will collectively have a better particle-based coverage of more of concept space than any one mind. But critically, here the distributional coverage is emergent and not represented in any individual (cf. Sloman & Fernbach, 2017).

Explaining our limitations

As we noted at the outset, perhaps the most challenging thing the evolutionary perspective demands is humility about the quality or finality of our current ideas. It implies our belief system is radically "located" and worse, that the alternatives we actually conceive of are shaped by blind luck and limited to relatively minor, or "local" modifications (relative to our entire belief system). This gives us the sensation of inhabiting a well posed problem of selection: i.e., Which of the alternatives that we have in mind is the most promising? But it also necessitates that beyond our conceptual horizon, we discover better alternatives only through what we might call "concerted randomness" as in the analogy of prospecting for gold. We think this perspective helps explain a number of the core patterns of biased or suboptimal behavior found across behavioral experiments in cognitive psychology. We here simply highlight how this proposal anticipates anchoring, probability matching and confirmation bias.

Anchoring. Human judgments are often "anchored" to existing values in seemingly arbitrary and non-optimal ways. On the evolutionary view, this is not at all surprising. Anchoring is an almost inevitable byproduct of any form of local search in an open possibly multimodal space of possibilities. On our view, any inductively established expertise—from domain general priors to learning mechanisms (both built in and learned during development)—are subject to the same issue. They are all incrementally acquired

and therefore non-independent samples from a latent theoretical posterior. This makes them potentially local optima, no different from the suboptimal products of biological evolution such as the giraffe's 5 meter laranyeal nerve, forced to travel around the aorta due to evolutionary selection that pre-dates the elongation of giraffes' necks (Wedel, 2011).

For agents who cannot try all possible hypotheses in parallel, making tweaks to their current beliefs is usually a better bet than starting over. This is because variants of one's current hypothesis can, in general, be expected to have higher fitness than a completely random guess (such as a new conceptual system sampled wholesale from an expressive and untuned PCFG). Understanding anchoring as a consequence of local sampling and an established library of concepts explains why the phenomenon is so ubiquitous (Bramley et al., 2017; Lieder et al., 2018), but also predicts the circumstances under which the effect can be reversed such that people's judgments are biased away from a provided anchor. Spicer, Zhu, Chater, and Sanborn (2022) show this can happen when an anchor is sufficiently close to the learner's initial hypothesis that their proposal mechanism can tend to push them away from it on average. Generally, we think this also lines up with the idea that mature cognition is characterized by the increasingly rigidity of its established concepts and difficulty in entertaining foundational changes (Gopnik et al., 2017).

Probability matching: Another common finding in human decision making is that people tend to select options in proportion to their posterior probability of being the most valuable, as opposed to reliably selecting the best option, in an apparent violation of rational decision theory (Shanks, Tunney, & McCarthy, 2002). One explanation for this is as a kind of solution to the explore–exploit tensions in action selection (Sajid, Ball, Parr, & Friston, 2021; Thompson, 1933). However, we think it is also relevant to note that probability matching is also the best-case-in-expectation for a learner adopting the endpoint of a single local search chain (see Bramley et al., 2017; Vul, Goodman, Griffiths, & Tenenbaum, 2014). To the extent that one has space to generate and compare multiple alternative possibilities, one can start to maximize over these options rather than simply

match as in an MCMC search. But, as we have already argued, selecting among a alternatives is something we can only do "locally". If we repeatedly locally mutate and maximize we are liable to become trapped in a local optimum in the long run, as happens with gradient descent algorithms. When it comes to revising our entire world model it would seem we simply do not have the space in our minds to entertain wholesale alternatives, nor the capability to generate alternatives far enough apart or numerous enough to allow for any kind of "distributional" coverage of the posterior that would allow for selection of its maximum a posteriori.

Confirmation bias. A third well known phenomenon is the tendency of learners to perform experiments or seek out data partial to their current favored hypothesis over evidence that would distinguish maximally among the full space of possibilities (Klayman & Ha, 1987; Nickerson, 1998). In a classic demonstration, Wason (1960) asked participants to identify a hidden rule and initially simply told them that the sequence 2–4–6 followed the rule. Wason's intended true rule was simply "ascending numbers" but participants rarely came up with this. In fact they reasonably think of hypotheses more likely to produce the sequence such as "triplets increasing by 2", or "consecutive even numbers". In testing these, they tended to get only confirmatory feedback (i.e. the outcome was consistent with their hypothesis) and so failed to narrow in effectively on the target concept. Again, we think this makes sense under the evolutionary perspective. Since the space of concepts is infinite and learners can only base their exploration on distinguishing among hypotheses they have actually generated for consideration (or else behave randomly). This means that when it comes to learning actively in an infinite hypothesis space of world models, any hypothesis-driven strategy will appear confirmatory from a "god's eye view" of the full possibility space.

Is cognition really blind?

On the face of it, characterizing cognition as a 'blind watchmaker' (Dawkins, 1986) operating without design is quite a strange idea—as the human mind is surely the

paradigmatic 'designer'. We agree the idea feels counterintuitive. This could be partly because idea generation is an aspect of cognition has resisted formalization and which we associate habitually with conscious control and free will, making it particularly strange to describe in algorithmic terms.[7] However, we think this applies to any mechanistic account of cognition, so does not pick out what feel strange about this claim in particular. More specifically, we suspect it may stem from a tendency in the rational analysis tradition to combine explorative and exploitative components of inductive reasoning by treating them as all part of a deterministic normative model. Probabilistic models of cognition (the tradition we broadly ascribe to) succeed in describing how our thoughts and behavior are highly directed in ways that reflect expectations we have built through experience. This is how actual watchmakers make watches efficiently, avoiding repeating the tedious blind walk of trial-and-error by caching and resusing techniques that seem to work. However watchmakers have to be able to develop and improve on their skills and occasionally improve on the cutting edge in watch design. Under a normative analysis, this part of the learning problem is easy to neglect because the hypothesis space is laid out in the idealization of the problem.

It feels natural to give credit to someone for inventing a better watch and seems reasonable even to say that they might set out deliberately to do so. Indeed, it seems more likely that an expert could improve on the state of the art in watchmaking than an amateur. But it also seems natural to distribute that credit, in final analysis, between features wholly compatible with our claims in this paper: (1) preexisting expertise (allowing them to focus "near" to promising watch designs, or target their known shortcomings), (2) perseverance (in trying many variations out) and (3) luck (in landing on something worth keeping). In this sense, the success of such an endeavor could be

---

[7] Incidentally, free will in the lay-sense articulated by John Locke of performing behaviors without a physical cause would amount to behaving randomly under a mechanistic theory of cognition (Dennett, 2015).

newsworthy *because* it is not guaranteed, which seems to fit nicely with the metaphor of innovation in cognition operating via local blind variation.

Conclusions

We have proposed that minds develop their inner world models through mechanisms of blind local variation and selection. We showed how the learning as program induction framework helps make this counterintuitive idea concrete and lends several algorithms as potential process accounts. A little differently to other proposals in this area, we have tried to motivate that the key explanatory virtue of these algorithms is not that they are, or can be, calibrated to respect principles of normative inference. Rather, we argued that it is because they produce variation recursively, while providing a receptacle (a current hypothesis and/or current concept library) allowing for the selective retention of the more valuable or promising products of this variation. We suggested that this can capture how it is that a hierarchical generative world model can grow within a mind through the recursive composition of a handful of basic operations. Overall, we suggest that this "algorithm-level" view is important for understanding cognition, in particular explaining why our thoughts are fundamentally anchored, order dependent and unpredictable even to ourselves.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*.

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Press.

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., ... De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, *29*.

Ashby, W. R. (1952). Can a mechanical chess-player outplay its designer? *The British Journal for the Philosophy of Science*, *3*(9), 44–57.

Aunger, R. (2000). Darwinizing culture: The status of memetics as a science.

Badcock, P. B., Friston, K. J., & Ramstead, M. J. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*, *31*, 104–121.

Borges, J. L. (1941/1998). The library of babel. *Collected fictions*.

Bourgin, D., Abbott, J., Griffiths, T., Smith, K., & Vul, E. (2014). Empirical evidence for markov chain monte carlo in memory search. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (Vol. 36).

Bowlby, J. (1969). *Attachment and loss*. New York: Basic Books.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38.

Bramley, N. R., Jones, A., Gureckis, T. M., & Ruggeri, A. (2022). Children's failure to control variables may reflect adaptive decision-making. *Psychonomic Bulletin & Review*, *29*(6), 2314–2324.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Bramley, N. R., Rothe, A., Tenenbaum, J. B., Xu, F., & Gureckis, T. M. (2018). Grounding compositional hypothesis generation in specific instances. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Bramley, N. R., & Ruggeri, A. (2022). Children's active physical learning is as effective and goal-targeted as adults'. *Developmental Psychology*.

Bramley, N. R., Schulz, E., Xu, F., & Tenenbaum, J. (2018). Learning as program induction.

Bramley, N. R., & Xu, F. (2023). Active inductive inference in children and adults: A constructivist perspective. *Cognition*, *238*(105471).

Brooks, S. (1998). Markov chain monte carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)*, *47*(1), 69–100.

Buchanan, D., Tenenbaum, J., & Sobel, D. (2010). Edge replacement and nonindependence in causation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32).

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review*, *67*(6), 380.

Campbell, J. O. (2016). Universal darwinism as a process of bayesian inference. *Frontiers in Systems Neuroscience*, 49.

Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59–68.

Carroll, S. B. (2005). *Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom* (No. 54). WW Norton & Company.

Castillo, L., León-Villagrá, P., Chater, N., & Sanborn, A. N. (2023). Explaining the flaws in human random generation as local sampling with momentum.

Changeux, J.-P. (1997). *Neuronal man: The biology of mind* (Vol. 21). Princeton

University Press.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive science*, *37*(6), 1171–1191.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811–823.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. N. (2020). Probabilistic biases meet the bayesian brain. *Current Directions in Psychological Science*, *29*(5), 506–512.

Chomsky, N. (1959). *Chomsky, n. 1959. a review of bf skinner's verbal behavior. language, 35 (1), 26–58.* JSTOR.

Church, A. (1963). Application of recursive arithmetic to the problem of circuit synthesis. *Journal of Symbolic Logic*, *28*(4).

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, *10*(8), 370–374.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Conant, R. C., & Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

Darwin, C. (1859/2004). *On the origin of species, 1859*. Routledge.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412.

Davis, Z. J., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*(5), e12839.

Daw, N., & Courville, A. (2008). The pigeon as particle filter. *Advances in Neural Information Processing Systems*, *20*, 369–376.

Dawkins, R. (1976). *The selfish gene*. Oxford University Press.

Dawkins, R. (1982). Replicators and vehicles. *Current problems in sociobiology*, *45*, 64.

Dawkins, R. (1983). Universal darwinism. *Evolution from molecules to men*, 403–425.

Dawkins, R. (1986). *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. WW Norton & Company.

Dechter, E., Malmaud, J., Adams, R. P., & Tenenbaum, J. B. (2013). Bootstrap learning via modular concept discovery. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*.

Dennett, D. C. (1995). Darwin's dangerous idea. *The Sciences*, *35*(3), 34–40.

Dennett, D. C. (2015). *Elbow room, new edition: The varieties of free will worth wanting*. mit Press.

Duhem, P. M. M. (1954). *The aim and structure of physical theory*. Princeton University Press.

Edelman, G. M. (1993). Neural darwinism: selection and reentrant signaling in higher brain function. *Neuron*, *10*(2), 115–125.

Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., Morales, L., Hewitt, L., ... Tenenbaum, J. B. (2021). Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd ACM Sigplan International Conference on Programming Language Design and Implementation* (pp. 835–850).

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*(2), 737–756.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to

explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.

Fodor, J. A. (1978). Tom swift and his procedural grandmother. *Cognition*, *6*(3), 229–247.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Fränken, J.-P., Lucas, C. G., Bramley, N. R., & Piantadosi, S. T. (2023). Modeling infant object perception as program induction. In *Proceedings of the 2023 Computational Cognitive Neuroscience Meeting.*

Fränken, J.-P., Theodoropoulos, N. C., & Bramley, N. R. (2022). Algorithms for adaptation in inductive inference. *Cognitive Psychology*.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gershman, S. J. (2017). On the blessing of abstraction. *The Quarterly Journal of Experimental Psychology*, *70*(3), 361–365.

Gershman, S. J. (2019a). How to never be wrong. *Psychonomic Bulletin & Review*, *26*, 13–28.

Gershman, S. J. (2019b). What does the free energy principle tell us about the brain? *arXiv preprint arXiv:1901.07945*.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In *Oxford handbook of causal reasoning.*

Ginsburg, S. (1966). *The mathematical theory of context free languages.* McGraw-Hill Book Company.

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.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661.

Guez, A., Silver, D., & Dayan, P. (2012). Efficient bayes-adaptive reinforcement learning using sample-based search. *Advances in Neural Information Processing Systems*, *25*.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2022). From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*.

Hemmer, P., & Steyvers, M. (2009). A bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Hodgson, G. M. (2005). Generalizing darwinism to social evolution: Some early attempts. *Journal of Economic Issues*, *39*(4), 899–914.

Hohwy, J. (2013). *The predictive mind*. OUP Oxford.

Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought*. MIT press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: the Bayesian approach*. Open Court Publishing.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161–1166.

Johnson, M., Griffiths, T., & Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in Neural*

Information Processing Systems, *19*.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–1243.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

Knight, F. H. (1921). Risk, uncertainty and profit, hart. *Schaffner & Marx*.

Kounios, J., & Beeman, M. (2009). The aha! moment: The cognitive neuroscience of insight. *Current Directions in Psychological Science*, *18*(4), 210–216.

Kripke, S. A. (1980). Naming and necessity: Lectures given to the princeton university philosophy colloquium. In *Semantics of natural language* (pp. 253–355). Springer.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Liang, P., Jordan, M. I., & Klein, D. (2010). Learning programs: A hierarchical bayesian approach. In *ICML* (pp. 639–646).

Liang, Y., Tenenbaum, J., Le, T. A., et al. (2022). Drawing out of distribution with neuro-symbolic generative models. *Advances in Neural Information Processing Systems*, *35*, 15244–15254.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*, 322–349.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147.

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing bayesian transfer. *Visual Neuroscience*, *26*(1), 147–155.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic

concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.

Marr, D. (1982). *Vision*. New York: Freeman & Co.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Michalski, R. S. (1969). On the quasi-minimal solution of the general covering problem.

Morgan, T. J., Suchow, J. W., & Griffiths, T. L. (2022). The experimental evolution of human culture: flexibility, fidelity and environmental instability. *Proceedings of the Royal Society B*, *289*(1986), 20221614.

Nelson, J. D. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

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

Perfors, A. (2012). Bayesian models of cognition: What's built in after all? *Philosophy Compass*, *7*(2), 127–138.

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines*, *31*, 1–58.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392.

Pinker, S. (2003). *How the mind works*. Penguin UK.

Plotkin, H. (1997). *Evolution in mind: an introduction to evolutionary psychology*. Harvard University Press.

Popper, K. R., et al. (1979). *Objective knowledge: An evolutionary approach* (Vol. 49).

Clarendon press Oxford.

Putnam, H. (1975). The meaning of" meaning".

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection.

Quine, W. v. O. (1969). *Word and object*. MIT press.

Quiroz, M., Kohn, R., Villani, M., & Tran, M.-N. (2018). Speeding up mcmc by efficient data subsampling. *Journal of the American Statistical Association*.

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster teaching via pomdp planning. *Cognitive Science*, *40*(6), 1290–1332.

Rafferty, A. N., LaMar, M. M., & Griffiths, T. L. (2015). Inferring learners' knowledge from their actions. *Cognitive Science*, *39*(3), 584–618.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, *26*(5), 521–562.

Rule, J. S., Schulz, E., Piantadosi, S. T., & Tenenbaum, J. B. (2018). Learning list concepts through program induction. *BioRxiv*, 321505.

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in Cognitive Sciences*, *24*(11), 900–915.

Rutherford, E. (1911). The scattering of α and β particles by matter and the structure of the atom. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *21*(125), 669–688.

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: demystified and compared. *Neural Computation*, *33*(3), 674–712.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144.

Sanborn, A. N., Zhu, J., Spicer, J., Leon-Villagra, P., Castillo, L., Falben, J., ... Chater, N. (2022). Noise in cognition: Bug or feature?

Schönfinkel, M. (1924). Über die bausteine der mathematischen logik. *Mathematische*

Annalen(92), 305–316.

Settles, B. (2009). Active learning.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Sharot, T. (2011). The optimism bias. *Current biology*, *21*(23), R941–R945.

Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. Riverhead Books.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Spicer, J., Zhu, J.-Q., Chater, N., & Sanborn, A. N. (2022). Perceptual and cognitive judgments show both anchoring and repulsion. *Psychological Science*, *33*(9), 1395–1407.

Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, *10*(2), 99–127.

Suchow, J. W., Bourgin, D. D., & Griffiths, T. L. (2017). Evolution in mind: Evolutionary dynamics, cognitive processes, and bayesian inference. *Trends in Cognitive Sciences*, *21*(7), 522–530.

Summers, P. D. (1977). A methodology for lisp program construction from examples. *Journal of the ACM (JACM)*, *24*(1), 161–175.

Sydow, M. v. (2012). *From darwinian metaphysics towards understanding the evolution of evolutionary mechanisms-a historical and philosophical analysis of gene-darwinism and universal darwinism*. Universitätsverlag Göttingen.

Szollosi, A., & Newell, B. R. (2020). People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences*, *24*(12), 1008–1018.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Thaker, P., Tenenbaum, J. B., & Gershman, S. J. (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology*, *77*, 10–20.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3-4), 285–294.

Tian, L., Ellis, K., Kryven, M., & Tenenbaum, J. (2020). Learning abstract structure for drawing by efficient motor program induction. *Advances in Neural Information Processing Systems*, *33*, 2686–2697.

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology*, *11*(4-5), 375–424.

Tooby, J., & DeVore, I. (1987). The reconstruction of hominid behavioral evolution through strategic modeling. *The evolution of human behavior: Primate models*, 183–237.

Turing, A. M., et al. (1936). On computable numbers, with an application to the entscheidungsproblem. *Journal of Math*, *58*(345-363), 5.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82.

Vélez, N., Christian, B., Hardy, M., Thompson, B. D., & Griffiths, T. L. (2023). How do humans overcome individual computational limitations by working together? *Cognitive Science*, *47*(1), e13232.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Wedel, M. J. (2011). A monument of inefficiency: The presumed course of the recurrent laryngeal nerve in sauropod dinosaurs. *Acta Palaeontologica Polonica*, *57*(2),

251–256.

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 681–688).

Williams, G. C. (1966). *Adaptation and natural selection: A critique of some current evolutionary thought* (Vol. 61). Princeton University Press.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Zhao, B., Bramley, N. R., & Lucas, C. (2022). Powering up causal generalization: A model of human conceptual bootstrapping with adaptor grammars. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Zhao, B., Lucas, C. G., & Bramley, N. R. (2022). How do people generalize causal relations over objects? A non-parametric Bayesian account. *Computational Brain & Behavior*, *5*(1), 22–44.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2023). A model of conceptual bootstrapping in human cognition. *Nature Human Behavior*.