Available online at www.sciencedirect.com

ScienceDirect

Physics of Life Reviews 11 (2014) 329–364

PHYSICS of LIFE reviews

www.elsevier.com/locate/plrev

Review

Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition

W. Tecumseh Fitch

Dept. of Cognitive Biology, University of Vienna, 14 Althanstrasse, Vienna, Austria

Received 5 February 2014; accepted 9 March 2014

Available online 13 May 2014

Communicated by L. Perlovsky

Abstract

Progress in understanding cognition requires a quantitative, theoretical framework, grounded in the other natural sciences and able to bridge between implementational, algorithmic and computational levels of explanation. I review recent results in neuroscience and cognitive biology that, when combined, provide key components of such an improved conceptual framework for contemporary cognitive science. Starting at the neuronal level, I first discuss the contemporary realization that single neurons are powerful tree-shaped computers, which implies a reorientation of computational models of learning and plasticity to a lower, cellular, level. I then turn to predictive systems theory (predictive coding and prediction-based learning) which provides a powerful formal framework for understanding brain function at a more global level. Although most formal models concerning predictive coding are framed in associationist terms, I argue that modern data necessitate a reinterpretation of such models in cognitive terms: as model-based predictive systems. Finally, I review the role of the theory of computation and formal language theory in the recent explosion of comparative biological research attempting to isolate and explore how different species differ in their cognitive capacities. Experiments to date strongly suggest that there is an important difference between humans and most other species, best characterized cognitively as a propensity by our species to infer tree structures from sequential data. Computationally, this capacity entails generative capacities above the regular (finite-state) level; implementationally, it requires some neural equivalent of a push-down stack. I dub this unusual human propensity "dendrophilia", and make a number of concrete suggestions about how such a system may be implemented in the human brain, about how and why it evolved, and what this implies for models of language acquisition. I conclude that, although much remains to be done, a neurally-grounded framework for theoretical cognitive science is within reach that can move beyond polarized debates and provide a more adequate theoretical future for cognitive biology.

Keywords: Cognitive science; Comparative cognition; Computational neuroscience; Neurolinguistics; Mathematical psychology; Formal language theory

http://dx.doi.org/10.1016/j.plrev.2014.04.005

1. Introduction: the central role for explicit theory in contemporary cognitive science

Every scientific breakthrough has its seed in some scientist's intuition, often couched as a metaphor or analogy: "heat is like a fluid", "electrical energy is like a wave" or "light is made of particles". Such vivid and intuitive conceptions are central to our initial conceptions of the physical world, and historically have played an important role in developing rigorous formal models. Typically, research progress involved testing the predictions from at least two plausible formalized models (e.g. "wave" versus "particle" theories of light).

But in physics today, it is the empirically-vetted formal models – the result of iterated testing of multiple plausible hypotheses – that are considered to represent ground truth. In modern science, Maxwell's equations show us how electromagnetic energy propagates, and the question of whether light is "really" a particle or a wave seems almost childish. Once a body of successful theory is in place, scientists are expected to update or re-educate their intuitions. This was not always the case: for physicists in the era of Newton or Descartes, intuitive conceptions played a more central role, and the mathematics simply provided useful accessories to that core. But today, many fundamental principles governing the physical world are well-understood, and such non-intuitive conceptions as "light behaves like both particles and waves", "heat is the movement of molecules", the relativistic equivalence of energy and mass, or the uncertainty principle in quantum mechanics are taken for granted as truths to which scientists' intuitions must simply adapt [50,83]. This is the insight at the heart of Mermin's famous dictum regarding quantum mechanics: "shut up and calculate!" [147]. This primacy of formal, mathematical models over pre-scientific intuitive conceptions and analogies is perhaps the clearest indication that physics is a mature science. A maturing cognitive science must eventually undergo a similar conceptual transition, acknowledging the primacy of formal and quantitative models over the metaphoric and intuitive "folk psychological" models that still dominate our discipline [35].

1.1. Metaphoric versus formal computational models in the cognitive sciences

There is of course nothing new in my suggestion that the cognitive sciences should embrace computational theory and formal models (e.g., [4,171]). What is novel in this paper is the specific framework I will advocate, which in a nutshell is cellular, computational and predictive, and the perspective I adopt, which is biological and comparative (treating empirically-evaluated similarities and differences between species as crucial sources of information for cognitive science). I will argue that success in this formalization enterprise will require triangulation between three disciplines: neuroscience (providing the firm physical foundations of brain function), the cognitive sciences (including cognitive and mathematical psychology, linguistics and musicology) and cognitive biology (providing a comparative viewpoint), and that the bridging functions between all three domains should build upon the insights of computer science (theories of computability and algorithms).

My point of departure will be Marr's [140] tri-partite approach to vision (Fig. 1A), which advocates separate specifications of the problem (what Marr terms "computation"), of the problem-solving approach ("algorithm"), and of algorithmic instantiation in neural circuitry ("implementation"). But rather than mapping the three disciplines onto Marr's hierarchy of computation/algorithm/implementation, I will suggest that *each* interdisciplinary bridge must incorporate explanations at *each* Marrian level (cf. [184]). My notion of "computation" is thus more fundamental, and more pervasive, than Marr's, and goes beyond simply "stating the problem" to incorporate all that modern computer science has to teach about which problems are simple, challenging, or intractable, and about what classes of algorithms can solve each type. Thus, although the framework outlined here accepts the importance of Marr's explanatory levels as originally described, I call for a revision in the manner in which they are used (cf. [172]), and I argue against the now standard separation of computational accounts from implementational ones (Fig. 1B). Instead, the insights of modern computer science permeate all three of Marr's levels, and his conception of "computational theory" was too narrow and confined to the most abstract level.

The general structure of the computational framework I advocate here is shown in Fig. 1C, which illustrates the need for productive interdisciplinary bridges between three disciplines, to form the vertices of a "computational triangulation" approach. The first two vertices are no surprise: the neurosciences and the cognitive sciences. Building bridges between these two domains is the traditional "grand challenge" of cognitive science of Fig. 1B, to bridge the apparent gap between brain and mind. However, the third, comparative, vertex traditionally plays a minor role in cognitive science. But to fully leverage the importance of algorithmic and implementational insights in constraining cognitive theories, species differences in cognition need to be identified and understood. I suggest that modern biol-

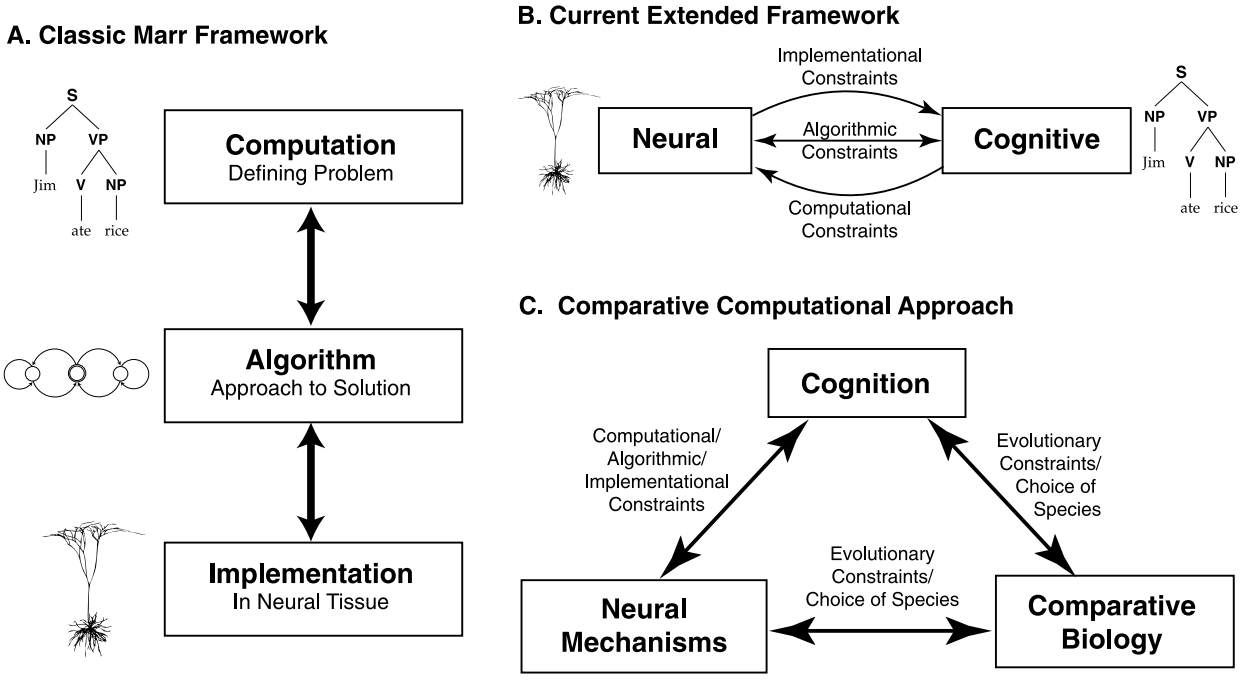

Fig. 1. Marr's Three Levels Panel A), contrasted with the approach developed in the current paper (B and C). See text for details.

ogy should play a central role this overall bridge-building enterprise, not least because of the impressive progress in comparative neuroscience and animal cognition made in the last few decades.

Another important point is suggested in Fig. 1B, illustrating an issue that should play a more central role in formalizing cognition: the concept of **tree-based computation**. Wherever one looks in the cognitive sciences, from the actual structure of single neurons right up to the most complex abstract structures of linguistics, key data structures or computational elements have a tree-shaped form. I suggest that this is not accidental, but rather reflects several computational virtues of tree-based computation (at both algorithmic and data structure levels) that are well-known to computer scientists, and have been independently "discovered" by natural selection. Some branches of cognitive science, like linguistics, take the centrality of trees and hierarchical structure for granted, and the tree-like structure of neurons is simply a fact of neuroscience, but in many branches of cognitive science, the conceptual importance of trees remains a topic of debate (e.g. in cognitive musicology cf. [122]). But there is considerable evidence (cf. [98,126,141,185]) that trees are implicit in many other aspects of (human) psychology, including motor planning, categorization, working memory (via "chunking") and perception (via the "local/global" distinction). Explicitly recognizing and formalizing this aspect of cognition will enable deeper connections to be discovered between these domains.

Finally, I argue that progress in cognitive biology will require us to discard certain intuitive oppositions that still play dominating (and I think destructive) roles in contemporary debates. Most prominently, the dichotomy between associationist versus cognitive models (or "statistics versus rules"), along with the "learned" versus "innate" dichotomy, are pre-scientific intuitive oppositions that are ill-defined, and thus ultimately untestable, and pernicious to empirical inquiry. I will argue that they need to be replaced with testable, formalized hypotheses drawn from a theory of biological computation, which can provide an overarching framework within which both individual learning and biologically-constrained species differences play important roles, and within which the associationist/cognitive dichotomy simply dissolves away. Thriving key components of such a formal framework already exist in computational neuroscience, mathematical psychology, in what I will call "predictive systems theory", and are highlighted below.

1.2. The current role of formalization in the cognitive sciences

The current formalization situation in the cognitive sciences today is quite different from that in physics or chemistry. Core folk-psychological concepts remain central (memory, learning, mental representations, categories…).

These are typically defined heuristically or introspectively, and research proceeds by breaking them down into sub-concepts (e.g. working, declarative, procedural and episodic memory) which are treated as primitives. Theories are often defined visually by connecting these primitive boxes together in diagrams representing the flow of information and control. Mainstream cognitive neuroscience often attempts to map such presumed primitives onto brain circuitry at a macroscopic level (e.g. highlighting the role of the hippocampus in memory formation or the basal ganglia in procedural memory), without explicit formalization of either side of this mapping. This approach is sometimes derided as "boxology" [19] or "neo-phrenology" [73], because it lacks the formal mathematical foundations of cell-level computational neuroscience or mathematical psychology. Nonetheless, a very considerable amount of empirical knowledge and understanding has accumulated in the context of such macro-level cognitive concepts as working memory or exemplar-based categorization, and it seems clear that any successful formal theory of cognition must be compatible with the constraints flowing from these cognitive data.

Of course, most cognitive scientists agree on the desirability of a body of formalized theory capable of linking well-understood properties of individual neurons to computational properties of neural circuits, and from there to existing cognitive models. Cognitive biologists would add that this enterprise needs to include explanations of why species differ, while acknowledging the many cognitive and computational similarities among species. But there are many links missing at present from a biologically-grounded theory of cognition and, despite pockets of formalization, bridging theories between well-formalized models remain rare (with some exceptions described below).

1.3. The role of computer science in the bridging enterprise

One important exception to this broad generalization about the cognitive sciences comes from computer science and the theory of computation. Early pioneers like Turing, Church and Post devised very different intuitive models of computation, and formalized them in very different ways, but by 1936 it became clear that these are all formally equivalent in the sense that they can represent precisely the same class of possible computations [38,39,216]. These different intuitive approaches are thus all equally valid. For historical reasons the Turing machine has become the placeholder for all of these models, although each formal approach is recognized as useful in different domains (e.g. Church's lambda calculus in semantics, or Post's rewrite systems in computational linguistics). Practically, computer scientists recognize that most existing programming languages are Turing equivalent. Thus, the choice between them is a matter of taste and practicality rather than objective superiority: one's choice of Python over C++ or Java is considered a topic for bar-room banter rather than pitched battles or scholarly debate. Much of the work in contemporary computer science involves distinguishing between hard and easy problems, devising algorithms that efficiently solve them, or showing why particular approaches do not scale successfully. Practical considerations play a central role, but not to the exclusion of theoretical rigor. Based on this, I suggest that computer science, and the theory of computation more generally, represents an important pocket of scientific maturity relative to the other cognitive sciences.

1.4. What about the autonomy of cognition from (neural) implementation?

Some cautious cognitive scientists or philosophers might find it naïve or even hubristic to attempt formalization and "neural reduction" of cognition, given our current incomplete understanding. Why formalize, one might ask, before we even have a clear intuitive idea about what the relevant entities in our theories might be? Furthermore, following Marr's classic division of vision into computational, algorithmic, and implementational components [140], many cognitive scientists argue for autonomy of the computational component from the messy details of implementation in the brain: the "autonomy thesis" [66,67]. From this widely-accepted viewpoint, bridging neuroscience and cognitive science is not just premature, but fundamentally wrong-headed.

I believe that such caution is unwarranted and that the autonomy thesis has outlived its usefulness. Regarding premature formalization, the history of science provides abundant evidence that science progresses faster when embracing crisp, clear models, even incorrect ones, than by accepting insightful but fuzzy metaphors that include partial truths, but are resistant to disproof. The very attempt to formalize can reveal blind spots or inconsistencies, and clarify the crucial choice points at which some aspects of a complex problem are foregrounded and others abstracted away. Furthermore, explicit formalization and contrasting of *multiple* plausible models provides a source of new experimental paradigms, and thus generates new experimental data. As George Box memorably put it "all models are wrong, but some are useful", and contrasting the predictions of multiple models is the surest way to gain insight into which are

usefully wrong [21]. In contrast, non-formal models often encourage a confirmational approach, in which a favored model is contrasted with some implausible null hypothesis, and minor discrepancies between theory and experiment lead to tweaking or elaboration of the existing model. Empirically, this often leads to ever-finer modifications of existing experimental paradigms, rather than inspiring new paradigms that generate potentially disconfirming data [233]. I argue that we are better off formalizing, and failing, than not formalizing at all.

Regarding neuroscience and the "autonomy thesis" the traditional notion that we can study cognition with little or no attention to its mechanistic, neural basis is misleading, both theoretically and practically. Theoretically, the thesis is often justified by characterizing cognition as a Turing machine, capable of any possible computation, and observing correctly that there are many ways to build such a machine (including, apparently, both silicon transistors and carbon-based cells). But human language, the capacity that has best been characterized at a computational level, is far from a Turing-complete system. Computational linguists currently agree that generating the full string set of any human language requires weakly context sensitive grammars, and no more [210], and we have no good reason to believe that vision, music or motor control require more than this. Such systems can be implemented in computational systems vastly less powerful than Turing machines. Practically, in models of computation below the Turing limit, factors such as implementation and real-time computation become central, and must be rigorously characterized.

Furthermore, the autonomy thesis stems from a time when computational neuroscience remained immature, and placed few important restrictions on cognitive models [66]. But a central message of the current review is that times have changed, and modern neuroscience now offers a very rich body of data and models, ripe for unification with more traditional cognitive frameworks. Unlike the founders of cognitive science 50 years ago, we are lucky to have both considerable understanding of neuronal physiology to constrain theorizing, and the powerful tools of modern computers at our disposal to crunch data and simulate models. It thus does not seem premature to attempt to create a comprehensive formal framework for cognitive science, including complex domains like language and music.

Given the central role that language and complex cognition played in the history of cognitive science [75,151], it is somewhat ironic that today, the most successful bridge between neuroscience and psychology is that between classical associative learning theory and neural predictive coding (e.g. in the midbrain dopaminergic system; see below). Classic "cognitive" topics like language or problem solving are less clearly formalized, and formal models that make testable predictions at the neuronal level remain rare. Perhaps this is not surprising given the traditional acceptance of the autonomy thesis by many cognitive scientists, and the daunting complexity of language. Unfortunately, being based largely on invasive studies in animal "models", associative learning theory has had little place for species *differences* as a relevant aspect of the research programme. But if we hope to understand the fundamental differences between species, including those differentiating humans from other animals, we need a framework that allows us to conceptualize and understand such differences, with language as a key example.

1.5. Toward a computational framework for comparative cognition

Here I will proceed by identifying some relatively uncontroversial well-formalized principles of neural computation, and then providing suggestions for how these can be linked to classic cognitive metaphors via the higher-level abstractions of the theory of computation. I will rather ambitiously choose the cognitive domains of human language and music as the high-level cognitive domains to be explained, because linguistics and musicology provide us with relatively clear formalized models, in comparison to many other aspects of cognition [103,119,174].

The goal is to build bridging theories between cell-level computation (a concept which is itself rather new), computational principles of multi-cell circuits, and the high-level concepts of cognitive science (perception, memory, learning, categorization, language, music). Although I will make use of David Marr's three-way split between explanatory levels, I will extend Marr's vision-based framework in three ways. First, as recently advocated by [172], I will include comparative (evolutionary) issues as a central component of cognitive science (this is the "cognitive biology" aspect of the framework). Second, I will not align Marr's levels with disciplines (e.g. "neuroscience = implementation" or "cognition = computation"), but rather argue for applying all three levels within each discipline, suggesting that a *computational* account at the neural level is a necessity for understanding *implementational* issues at the cognitive level. Third, taking my lead from contemporary work in computational linguistics and the theory of algorithms, I will highlight the importance of trees as a basic data structure for understanding cognition, at least in humans. The framing of a wide variety of computational problems in terms of trees (e.g., binary search trees, decision trees, suffix trees, spanning trees, etc.) and the existence of efficient tree-processing algorithms to solve such problems

is well-known to computer scientists [162,207]. But the scope and power of tree-based computing remains less widely appreciated in the cognitive sciences, where sequences, matrices, and arbitrary graphs receive much more intensive attention.

At the implementational level, any biologically-grounded theory of cognition will need to accept important differences between neural wetware and contemporary computer hardware. Neurons are living cells – complex self-modifying arrangements of living matter – while silicon transistors are etched and fixed. This means that applying the "software/hardware" distinction to the nervous system is misleading. The fact that neurons change their form, and that such change is at the heart of learning and plasticity, makes the term "neural hardware" particularly inappropriate. The mind is not a program running on the hardware of the brain. The mind is constituted by the ever-changing living tissue of the brain, made up of a class of complex cells, each one different in ways that matter, and that are specialized to process information.

At the algorithmic level, it is now a commonplace that many of the cognitive problems successfully dealt with by organisms every day are, technically, computationally intractable: long is the list of NP-hard computational problems that organisms solve without obvious effort or delay [184,218]. The reason, of course, is that organisms make do with heuristic solutions, crafted by millions of years of evolution, that are "probably approximately correct" and which evade the worst-case scenarios that are the bane of computer algorithm designers [81,112,222]. Given that organisms must make use of such rough-and-ready algorithms, and that a body of theory exists to understand them, it would be perverse for cognitive scientists to ignore this important source of algorithmic constraints on computation-level theories.

At the computational level, many important insights of the connectionist era, particularly the core insights that neural computation is parallel and distributed, must be carried over into the future. But rather than seeing these insights in opposition to traditional "computational" approaches based on symbol processing, rules and representations, it seems increasingly clear that each of these approaches offers valuable insights, and they represent complementary rather than conflicting levels of explanation. I will suggest below that several computational insights from connectionist "neural network" models, especially the role of error in learning, in fact provide a better fit at the *cellular* level of computation than at the system level. The framework advanced below thus implies a redeployment of these insights down to the level of single-neuron computation.

1.6. Outline of what follows

In the rest of this paper, I will attempt to flesh out the proposed interdisciplinary framework sketched above. I start by pinpointing several traditional folk-psychological distinctions that have long framed debate in cognitive science, but are unhelpful because they really concern issues of emphasis, or focus on different levels of the theoretical framework, rather than reflecting clashing world views about how neural computation and cognition work (Section 2). These unhelpful but pervasive dichotomies include the long-running debate between nature versus nurture, and between connectionist/associationist approaches and "classical" cognitive models based on rules, representations, and programming analogies.

In Sections 3–6, I turn to the main (positive) messages of the paper. Section 3 argues that cognitive scientists need to take seriously the biological wetware within which cognitive computations are performed, especially the fact that **individual cells are the basic computational unit**, themselves constituting complex computing devices. This is a fact of modern neuroscience: neurons themselves do computation that cannot be captured as a simple "integrate and fire" node in a "neural" network. This fact implies that neither connectionist networks nor classical cognitive models are searching for the bridges to computational neuroscience in the right places. Section 4 emphasizes a view of brains as **predictive systems**, rather than representational engines, and the importance of this viewpoint in constraining our models of the algorithms we use to accomplish difficult tasks. Again the theory of computation, particularly algorithmic and coding efficiency, holds important lessons. Section 5 highlights the need to characterize species differences computationally. I suggest that at least for higher human cognitive abilities like language and music, the **ability to infer hierarchical (tree-based) structures is critical**. The theory of computation provides an explicit framework within which to characterize this capacity, and to probe the differing computational abilities of multiple species. Existing data collected within this framework suggests that representing explicit hierarchical structures is difficult, or impossible, for many non-human species. This appears to represent a qualitative difference among species, or at least a substantial quantitative difference with qualitative consequences for cognition.

In the 7th and final section, I will offer a more speculative account of the importance of tree-based computational approaches to cognition in understanding species differences and how they may be implemented in neural wetware. In the hope of mapping out some parts of the terrain that needs to be covered, I offer some speculative but testable hypotheses to spur research. My goal here is to exemplify rather than to exhaustively catalog: I do not mean to imply that the framework presented is limited to the domains discussed.

2. Initial diagnosis and proposed remedy: using formal models to move beyond intuitive distinctions

I first hope to diagnose what I see as a central problem in much contemporary cognitive research: an over-reliance on conceptual dichotomies that represent intuitive, ill-defined viewpoints rather than well-specified models or frameworks. My target dichotomies are "innate knowledge" versus "learning", on the one hand, and "associationist" versus "cognitive" models on the other. Neither of these distinctions are well-defined, and I will argue that both dichotomies have proven ultimately pernicious. Once we formalize the problems (e.g. using some variant of a Bayesian framework) it becomes clear that debates employing these traditional dichotomies foreground irrelevant and ultimately philosophical issues, while downplaying or ignoring the empirical issues of central relevance.

Throughout this essay, I will explore examples in which well-formalized and successful models reveal such traditional intuitive distinctions to be misleading. For example the recent successful bridging of formal learning theory (e.g. the Rescorla–Wagner rule) and midbrain dopaminergic reward prediction systems has revealed that no distinction is made at the neural level between "model-free" (Pavlovian) learning and "model-based" (cognitive) learning. Similarly, a Bayesian framework provides a natural home encompassing traditional probabilistic learning approaches, rule-based symbolic computation, and innate species differences (implemented as reliably-developing prior probability distributions), dissolving dichotomies that seem intuitively obvious from a folk-psychological viewpoint.

2.1. "Learned versus innate knowledge"

The so-called nature/nurture debate is one of the longest-running, and least productive, disputes in all of academic discourse. What makes its longevity remarkable is that virtually all informed commentators on the dichotomy agree that it is misleading and pernicious, representing what Patrick Bateson memorably dubbed "the corpse of a wearisome debate" [12]. Yet the nature/nurture opposition continues to spark strong feelings and polarize cognitive scientists (e.g., [11,169]).

There is no dearth of detailed analysis of this "nature *versus* nurture" scourge, from both philosophical [5,85] and biological [13,134] perspectives. There is a clear consensus that the term "innate" acts as a cover term for multiple distinct biological properties that, although typically conflated in folk psychology, need to be considered separately in modern science. For example, "innate" often connotes species- or domain-specificity to cognitive scientists. But many innate propensities (such as the satiating value of food for hungry individuals) are broadly shared by most animal species. Similarly, many "classical" computational mechanisms underlying vision (binocular disparity processing, shape from shading, etc.) are presumably both genetically-canalized and domain-specific, but are nonetheless shared among many mammals. Innately determined learning mechanisms, such as classical conditioning, are presumably both species- and domain-general. So "innate" does not necessarily mean "species specific" *or* "domain specific", and there is no justification for the unfortunate tendency to conflate genetic determination, species-specificity, and domain-specificity all together, or to misuse the term "innate" to refer to all of them [58].

An additional conceptual problem concerns **innate learning mechanisms**. To many psychologists, "innate" connotes that NO learning takes place, a precept psychologists often accept either implicitly (e.g., [178]) or explicitly (e.g., [189]). By accepting this disjunction, psychologists miss out on a clear and biologically-grounded escape path out of the nature/nurture quandary: the recognition of the existence and importance of "instincts to learn" [59,139] in understanding both learning and instinct. No biologist studying hunger would deny that animals learn about different foods, and no psychologist would deny that feeding, hunger, and digestion all have strong and reliably-developing inborn components. Most aspects of cognition involve "instincts to learn" – bundles of proclivities, biases and constraints that lead the organism to attend to particular aspects of situations, biasing them to form certain types of generalizations and not others [22,74]. The claim that human language or music have an innate component thus does not imply the (obviously fallacious) claim that no learning is involved in these domains. Because this issue has already been so thoroughly diagnosed, and antidotes suggested, I will not belabor these points further (cf. [5,85]): cognition in

humans and other animals is grounded in "instincts to learn," and to ask if a mature capacity is "innate" or "learned" is to immediately go astray in understanding it scientifically.

2.2. The Bayesian path beyond nature/nurture

The nature/nurture dichotomy provides my first illustration of the value of a formal approach, in this case of a Bayesian perspective, for moving beyond outworn intuitive metaphors. The Bayesian approach provides a formal mathematical framework within which organisms can update the parameters of internal models (θ), based on data (D). In particular, Bayes' Theorem shows that:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

which, if the data are already in hand reduces to the more compact and insightful form:

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

This says that our estimate of the probability of a model, given some data, is proportional to the product of the probability of this particular data, given the model $p(D|\theta)$, and our "prior" estimate of that model's probability, $p(\theta)$. Although this is a simple mathematical truth (a theorem), and the significance and power of this rule is widely appreciated among computer scientists or learning theorists, it is less common in classical rule-and-representation approaches to cognitive science. Indeed, such probabilistic, learning-based approaches are sometimes actively contrasted with symbolic approaches [11,168,197]. This is a shame, because the Bayesian formalism provides a natural home for both learning (data) *and* innate constraints on learning, via priors, [27].

I begin by observing that the folk psychological term "knowledge" can be formalized in terms of Bayesian priors on some probability distribution. Nativists and learning theorists will both agree that the response of an organism to a stimulus or a situation is influenced strongly by its (unconscious) estimation of the probabilities that different actions will lead to successful or unsuccessful outcomes (rewards or punishments). These probabilities can be represented and quantified as prior probabilities for those actions. This makes "where do such priors come from?" a central question. Some may be intuitively categorized as innate (for example, a newborn mammal's suckling response is a highly canalized reaction, present at birth), and some may intuitively appear learned (as for Pavlov's dogs salivating to a bell). But in both cases a biological substrate exists which provides the necessary prerequisites for learning.

In many cases the "learning" has been done on a phylogenetic time scale (through evolution): this is the province of reliably-developing "instincts", traditionally studied in classical ethology. In other cases a biologically-provided learning mechanism (such as classical conditioning) enables the organism to generate new priors in its own lifetime (the classical province of learning theorists). Even in the learning of wholly arbitrary stimulus/response associations, the computational mechanisms allowing the organism to compute statistics over stimuli, actions and consequences, along with appropriate recognition of "rewards" and "punishments", must be innate. And despite the appealing behaviorist dream that these evolutionary priors would turn out to reflect species-general "laws of learning", it has long been clear that species vary considerably in both what and how well they learn. Learning abilities vary across species [22,134], and individual organisms are biologically biased to form certain classes of associations and not others [74,200].

Notably, when formulated in probabilistic terms we can ask "what prior probabilities influence an organism's cognition and behavior in a particular circumstance", fully expecting the list of priors to include a mix of evolutionarily-acquired biases and others acquired by the individual organism over its ontogeny. This mix will vary for different types of knowledge and actions, and over the individual's lifespan. A baby does not "learn" to cry when hungry, or to laugh when tickled, but the child does learn that crying may influence caregiver behavior, just as the adult learns that crying or laughing are inappropriate in certain situations. There is no need to categorize either behavior as *either* innate or learned, as it will always be both, and indeed the answer will depend on the individual's age, experience and culture. Nor will insight come from treating this as a simple one-dimensional continuum, where traits are, for example, "60% learned and 40% innate": the probability space has many more dimensions than this simple continuum can encompass, and requires a much more nuanced and fine-grained analysis.

An important virtue of the Bayesian probabilistic approach to cognition is that it is, in an important sense, "substrate neutral". That is, after meeting certain minimal conditions about independence, statistical approaches can be applied

to, and probability distributions derived for, *any* definable type of entity or model. But the types of entities "visible" for probabilistic inference will vary from species to species. We obviously do not expect a color-blind organism to estimate probability distributions over colors, despite being fully equipped with perfect memory and a Bayesian inference engine. Central issues then become "what classes of stimuli can be recognized and counted?" and "what effects are rewarding?" About these biological questions, which are logically prior to probability estimation or application of Bayes' rule, the statistical framework remains initially silent: the answers must come from the history of that individual and the biology of its species.

2.3. Nature via nurture in language: the Bayesian perspective

A nice illustration of the value of both rules and statistics comes from computational linguistics, where a probabilistic approach over word- or letter-strings held sway for many years [186]. Applied to syntax, the probabilistic approach started by assuming that each word in a sentence predicts its immediate successor with some probability (represented by a two-word "bigram"). By calculating probability distributions over bigrams, sequentially, we can both generate sentences [198,199] and predict the probability of observed sentences. As Shannon noted, this approach fails for a relatively obvious reason: sometimes words are related to more distant words than their nearest neighbor. For example, in "the cat who scratched the dog ran away", we know that it was the cat who ran away, despite the presence in this word string of "the dog ran away". While we might attempt to get around this by using larger strings (n-grams, where trigrams are the most common), this will not solve the problem fully. Long-distance dependencies present a major problem for n-gram approaches both because there is no fixed distance between related words (so no given n will necessarily work), and because there is a massive combinatorial explosion of possible n-grams as n increases. Even for trigrams, the amount of data required to calculate the probabilities is vast, and even after processing a 38-million-word corpus, many specific trigrams known to be valid remain unattested [186].

Most contemporary approaches take a quite different approach, and attempt to estimate syntactic *structures* rather than word strings [27,110,163,165,186]. The problem remains statistical, but now the task is to estimate a tree structure S given a word string D (that is, $P(S|D)$). The capacity to generate a set of possible tree structures (via a context-free grammar) is thus at the heart of probabilistic computational linguistics [108,110,165]. But, just as a color-blind organism cannot do statistics over color, a Bayesian organism that happens to be incapable of perceiving or remembering tree structures cannot even begin to solve this problem. Considerable data suggests that many non-human organisms may be limited in precisely this way (see below). Thus the core issue is not whether an organism can learn statistical associations, or uses probability estimates to do so, but rather concerns what *types* of objects or stimuli it can process. From this perspective, a probabilistic Bayesian approach to language, and child language acquisition in particular, is compatible with the full spectrum of cognitive approaches, from nativist, generative approaches [109] to those that place domain-general learning systems at center stage [27]. Put in these terms, the questions become empirical, and include what types of models a particular organism is capable of generating, and to what degree its initial prior distribution over these models is uniform versus biased in one way or another.

While computational linguists already largely embrace this dichotomy-denying framework, both theoretical linguists and psycholinguists have been slow to follow, with some notable and important exceptions (e.g., [164,165]). Cognitive (neuro)science, in general, lags even further behind. Thus, the re-marriage of rules and statistics, and nature *via* nurture, is a revolution ready to occur.

2.4. "Cognitive" versus "associationist" interpretations of learning

The second long-running dichotomy I will criticize is particularly problematic in the animal cognition literature: a rigid distinction between "cognitive" and "associationist" explanations. "Associationist" explanations focus on stimulus-driven behavior, unconscious knowledge, and typically learning mechanisms thought to exist in a wide range of species (e.g. classical conditioning and associative learning). In contrast, "cognitive" explanations typically connote covert mental activity (going beyond perceptual stimuli and behavioral responses), as well as flexible context-dependence, goal-directed model-building, and often consciousness. Again, this basic dichotomy has deep roots (going back to the arguments between rationalist and empiricist philosophers) and continues to play a prominent role in contemporary cognitive science. As for the nature/nurture debate, the fact that this debate has remained unresolved for so long suggests that the central questions are not directly subject to empirical test or refutation. But again,

a clear and empirically successful body of formal theory exists which incorporates and subsumes the crucial elements of both perspectives: what I dub "predictive systems theory" (see below). This research field provides a beautiful illustration of the power of formalization to dissolve and defuse the intuitive folk-psychological distinction between "cognitive" and "associative" perspectives.

2.5. *Beyond the cognitive/associationist debate with predictive approaches*

I will highlight current work conceptualizing the dopaminergic reward system as a model-based predictive code, consistent with traditional Rescorla–Wagner learning theory [194]. Classical behaviorism eschewed all mentalistic "covert" variables in the analysis of behavior, and sought general laws of stimulus/response learning that would apply to all organisms and all stimuli. However, serious problems with these goals became evident, within the associative tradition, from demonstrations that certain types of associations can be learned much more easily than others. For example, a rat easily learns to associate a taste with later nausea, or a light flash with electric shock, but despite repeated pairings, does not learn to associate a light with nausea, or a taste with shock [74]. While these biases make sense from an evolutionary viewpoint, they presented a serious challenge to the notion of general, domain-independent "laws" of learning. An even more worrisome issue was presented by the phenomenon of "blocking" [180], where a previously-learned association (e.g. between a light and reward) "blocks" the acquisition of an equally-informative second association (e.g. between a sound and that same reward).

These and related phenomena led to a renaissance in learning theory in the 1970s, in which associative learning came to be seen in terms of **prediction** of rewards, rather than of simple stimulus-response linkages. While learning theorists still tended to characterize their views in associationist terms (e.g., [180]) it is important to recognize that in incorporating the organism's predictions about rewards, the theory took a very big step in a cognitive direction. For where do predictions come from? Clearly, from some internal model used by the organism to make predictions based on context and stimuli. Although not typically acknowledged, this requires some form of (perhaps unconscious) "mental" model, requiring learning theorists to "belatedly embrace the cognitive revolution" [40].

Again, a formal characterization helps to make the issues more explicit. Take the classic Rescorla–Wagner learning rule, probably the most successful and influential formal model in mathematical psychology:

$$\Delta V_x = \alpha_x \beta(\lambda - V_x)$$

where ΔV is the change in associative strength, lambda is the highest associative strength possible for the current reward, and V_x is the current predicted associative strength. Beta is a fixed parameter determining learning speed, and alpha_x is another rate parameter which indicates the "salience" of x to the organism in question. This equation can be expanded to compound cues x_1, x_2, x_n by stipulating that the total V is the sum of all V_xs.

The Rescorla–Wagner model is essentially just a simple error-minimizing model-fitting system, in which prediction accuracy is improved by continually evaluating the deviation between prediction and reality (the "prediction error"). The strength of the updating process is directly correlated with the "surprisingness" of the result [180,195]. The equation incorporates species-specificity via the two rate parameters, alpha and beta, where alpha indicates the salience of a particular sensory signal to the species and beta indicates an overall "general" learning speed. The reward value incorporated within the lambda term could also be species-specific (ants are rewarding to anteaters, as is clover to cows – but not vice versa).

More importantly, the Rescorla–Wagner model covertly entails a *cognitive* mechanism capable of generating the predicted outcome V_x, without which it would be impossible to calculate prediction error. Rescorla himself called the mechanism responsible for this a "representation" [180], though today it is typically couched in terms of a prediction based on a "model" [195]. Whatever we call this process, it is clearly cognitive in the sense of being based on a covert internal model, which is the key component that changes during learning. That is, although initially the prediction error will simply be directly proportional to the strength of the reward (lambda) because no prediction is made ($V_x = 0$), learning entails updating the system to the point where $V_x =$ lambda, at which point there is no remaining prediction error. Based only on this equation and the behavioral data, we might continue to debate just how "cognitive" such predictions are.

Fortunately, the prediction-based approach embodied in the Rescorla–Wagner equation has driven major empirical progress in understanding associative learning in neural terms [194–196], and leaves little doubt that the predictions driving associative learning are indeed cognitive (or as Schultz [195] puts it "model-based"). For example,

the firing rate of dopamine neurons in the mammalian midbrain appears to directly encode the prediction error term (lambda $- V$) from the Rescorla–Wagner model. But, crucially, the prediction error encoded by these neurons is based on *subjective* predictions rather than objective ones when the two conflict. For example, dopaminergic neurons react to the *subjective* reward value (discounted for time), rather than the raw amount of food, and encode *subjective* presence or absence of stimuli (rather than their actual physical presence or absence) when this is ambiguous [195].

This modern, neurally grounded interpretation of classical conditioning has come a long way from the reflexive analogy favored by Pavlov, or the orientation to the physical stimulus typical of behaviorists, moving in a clearly "cognitive" direction [180]. Nonetheless most contemporary learning theorists (or neuroscientists) would continue to characterize this process as "associative" and eschew cognitive, mentalistic terms in their own research. Furthermore, in the animal cognition literature, such "associative" explanations continue to be seen as alternatives to "cognitive" explanations, and often are considered to be *a priori* simpler and more parsimonious (cf. [23,24,93]).

This seems to me a clear case where careful, formalized models have fueled scientific progress and can help to build the needed bridges between psychology and neuroscience, and between theory and experiment. Observe that no victory has been declared in the long-running battle between cognition and association: instead, this traditional dichotomy is simply dissolving away and has lost much of its former significance (cf. [40,180,195]). Scientists working at this interface (e.g., [196]) do not appear overly concerned about whether their models are best characterized as "cognitive" or "associationist". Instead, they admirably focus on how well their models fit empirical data, and more importantly, generate testable predictions. The key point is that clear formalization of hypotheses and their predictions, combined with empirical investigation at the neural level, can lead to a dissolution of intuitive but fuzzy dichotomies. Again, the advance is being led by computationally-minded neuroscientists, and fields like cognitive psychology and animal cognition have some catching up to do.

The two examples above illustrate the promise of a formally grounded empirical approach to span the current gaps between the fields of cognition, neuroscience, and computation. With this critique behind us, I now turn to the more positive aspect of the paper. My goal is to map out a computational framework bridging theories between cellular neuroscience at the low end to computational perspectives on music and language at the high end.

3. Single-neuron computing: bridging cognitive theory and neuro-computational reality

The first important component of my framework concerns the computational power of single neurons (cf. [92,116,118,132]). It is a central insight of contemporary computational neuroscience, for at least the last decade, that traditional models of neural computation (from McCulloch–Pitts to recurrent connectionist networks) vastly underestimate the computational power of single neurons: It now appears that a single cortical pyramidal cell has roughly the power of a classical two-layer neural network [92]. The central importance of this fact is that it inserts a heretofore neglected level of biological agency – that of the single cell – as a key locus of learning and adaptation in computational models. Previously, learning from errors was modeled (probably unrealistically) as a process at the network-level, but we now know that error-based learning can take place in an individual cell, via essentially Hebbian processes like spike-timing dependent plasticity [125,132]. This re-estimation of neuronal computational powers also has some interesting implications for evolution, inserting the biology of living cells back into both computational modeling and philosophical debates about intentionality and consciousness (cf. [53]). But more importantly for the current argument, these neural facts necessitate a redeployment of some of the insights from nearly thirty years of work in connectionism [187], from the network down to the cell level.

3.1. The tradition: "consider a spherical neuron…"

There is an old joke about a theoretical physicist asked to help re-design a dairy farm who, after weeks of deep thought, begins his presentation with "Consider a spherical cow…" The joke is funny because it emphasizes the tendency of physicists to simplify a problem (often for ease of calculation) in ways that seem unrealistic, non-intuitive, or impractical. Perhaps less humorous is the fact that this has been precisely the standard approach to modeling neuronal function for roughly 50 years, starting with the very foundations of computational neuroscience [145]. This "spherical neuron" tradition exists despite the fact that the actual shape of most neurons is complex and treelike (see Fig. 2) – about as far from spherical as geometrically possible. Traditionally, this difference between model and reality

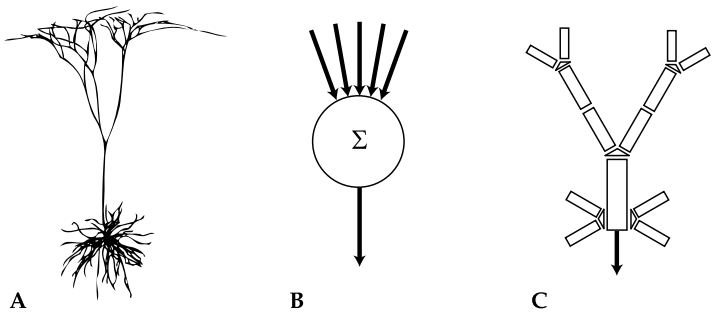

Fig. 2. A) A real neuron (cortical pyramidal cell), contrasted with B) the "spherical neuron" (traditional computing element, used in connectionist models) and a C) more modern multi-compartmental model of a neuron which acknowledges the computational complexity of single cells.

was seen as superficial and inconsequential, at least from a computational viewpoint. But in the last decade or so it has become clear that this is incorrect at a fundamental level. Nonetheless, even today, the vast majority of research in computational neuroscience and connectionist models involves networks of simple "spherical neurons" which simply sum and threshold their inputs (cf. [117,118,204]).

The traditional computational approach to the nervous system was first made rigorous by Warren McCulloch and Walter Pitts in a seminal paper demonstrating that one can translate any proposition couched in logical terms into specific binary networks of rather simple "units", proposed to be essentially equivalent to neurons [145]. These "neurons" simply sum all of their positive inputs (linearly) and apply a (nonlinear) threshold function; if the summed "activation" is above the threshold, the unit "fires" (generates a 1 instead of a 0). In the "classic" McCulloch/Pitts neuron, there were also inhibitory neurons, and if one or more impinging inhibitory neurons was active, the cell would not fire. The computationally crucial aspect of these units is that each implemented a nonlinear thresholding operation, which converts its essentially analog (real-valued) inputs into a discrete (binary or "logical") output. But all of the computational power in a McCulloch–Pitts system comes from the arrangement of these very simple units into networks. An aspect of this idealization, known to be oversimplified even at that time, was that a cell's current state depended on only its current inputs, and not its *own* previous state. This ignored the fact that, after a neuron fires, it goes through an intrinsic refractory period during which it cannot fire (or firing probability is reduced).

The demonstration of the computational power of networks had an immense effect on neuroscience because it rendered plausible the idea that biological neural circuits could be characterized in rigorous computational terms, and showed that the computations permitted were extremely powerful (cf. [153]). McCulloch and Pitts themselves claimed that their computational characterization was compatible with many different neurophysiological implementations, a claim that appeared to warrant abstracting away from cell-level details when investigating neural computation. Rightly or wrongly, this became a founding assumption of cognitive neuroscience, and ever since seemed to justify separating "cognition" from "cellular neurophysiology" (often misleadingly glossed as "software versus hardware"). Thus, after a cursory treatment of single-cell physiology, current cognitive neuroscience textbooks focus almost entirely on macroscopic neural circuitry, cortical regions or whole-brain function [76].

Despite its over-simplifications, the McCulloch–Pitts "neuron" became, and remains, a basic staple of computational neuroscience. This is particularly marked in the broad resurgence of connectionist networks (a.k.a. "neural networks" or "parallel-distributed processing" models) starting in the late 1980s and continuing today. By overcoming some computational limitations of earlier neurally-inspired models like Rosenblatt's Perceptron model [154], connectionists led a resurgence of "rule-free" associative models of the mind in cognitive science [44,180,187]. Although many of the tricks used in connectionist models are biologically unrealistic (e.g. back-propagation of error or "context units" in recurrent networks), these systems are provably very powerful [205], and have been used to successfully model various aspects of high-level computation including music and language [18,33,44]. However, research in the last decades has quite clearly demonstrated that models based on integrate-threshold-fire neurons are woefully inadequate as models of real nervous systems, for a simple and important reason: single neurons are much more powerful than McCulloch–Pitts cells or connectionist "units". These findings demonstrate both that there is nothing *neural* about "neural networks", and more importantly that cognitive scientists need to pay renewed attention to cellular neuroscience if we hope to understand the biological basis of cognitive computation and bridge the (apparent) mind/brain gap.

3.2. The modern view: computation and the single neuron

It has been obvious since Ramón y Cajal first published his lavish drawings of single neurons [175] that neurons are cells distinguished by their complex three-dimensional form, which at least in vertebrates typically closely resembles that of a tree (with axons and dendrites corresponding to leaf-bearing branches and the root system of a botanical tree). It long seemed possible that this cellular morphology simply represented a way of distributing the neuron's mass through a large volume of nervous tissues (e.g. to allow more connections with other neurons). The question of what, if any, *computational* significance the treelike form of neurons might have has only been conclusively resolved in the last decade [118]. We now know that, far from being passive summators, the different branches of the dendritic tree themselves act as computational elements (roughly equivalent to the "units" of a connectionist network) [92,132]. The presence of active (voltage-dependent) ion channels, as well as short-range interactions within dendritic branches, means that a single biological neuron is roughly equivalent in computational power to a two-layer neural network made up of hundreds of "units". It is now possible to simulate the complex function of a neuron using principles of cellular electrophysiology that are well-understood at a chemical and physical level [117], and ongoing advances in imaging and cell-level manipulation now make it possible to empirically observe such computations in living cells, and compare them to models (e.g., [26,80,221]). There can now be no doubt that the complex, tree-like form of individual neurons plays an important, and indeed central, role in determining the integration and firing behavior of cells, and thus in the computations of the networks they are part of. Furthermore, unlike silicon transistors, each neuron is different in form, in ways that reflect its past history and often matter computationally (cf. [53]).

At first blush this may seem like depressing news for cognitive scientists. As if the nervous system wasn't complex enough, we now need to worry about the complexity of single cells? But on closer consideration this increase in complexity at the local level may in fact enable us to simplify our theories at a more global level. Recognizing that single cells engage in complex computations allows us to incorporate a previously known aspect of biological agency – that of single cells – into our models of computation. Single cells respond to their environment by changing their form and composition, and free-living single-celled organisms can learn about their environment. By recognizing these capacities in neurons, which are after all eukaryotic cells inheriting two billion years worth of evolved adaptability, we can now reconceptualize the level at which certain types of learning (most notably Hebbian processes) take place. Things that made little physiological sense at a network level (e.g. back-propagation of error signals: [184]) make perfect biophysical sense in a cellular context, because when a cell fires an action potential, depolarization propagates back through the dendritic tree, providing a local "teaching" signal to each synapse. Furthermore, a host of well-investigated neurophysiological phenomena such as long-term depression or potentiation, mediated by NMDA (n-methyl-d-aspartate) receptors, provide a clear mechanism whereby a cell "learns" the required computations by updating its physical form (cf. [218]). Decades worth of computational insights derived from connectionist modeling can thus be redeployed at the cellular, rather than the network, level. The result – I believe – will be a major boost to attempts at bridging the cognitive/neural gap, building on the solid foundations of cellular biophysics and the fast-developing field of cellular computation.

3.3. Single-cell computing: further implications for cognitive science

There are several further implications of single-cell computation for cognitive science. The fact that computation at the cellular level occurs in trees (and not arbitrary, or fully connected, networks of synapses) should place important constraints on the kinds of elementary computations computable by a single cell, which in turn must influence the types of algorithms that can be efficiently realized in networks of tree-shaped neurons. This, in turn, should "percolate up" to the level of cognitive theories, allowing us to better evaluate which computational models and algorithms are biologically plausible at the implementational level. Thus, although it would be naïve to think that there is any direct connection between the tree form of individual neurons and the many abstract trees that populate the cognitive landscape, it is simply realistic to recognize that such abstract cognitive trees are implemented in computational machinery made up of tree networks. In terms of the "grand challenge" of cognitive neuroscience, explicitly building bridges from mind to brain, I suspect that having trees anchoring both sides of the bridge should help constrain and channel future bridge-building.

A second set of implications of single-neuron computation concerns species differences in cognition. We know that important species differences exist, from the perceptual (bats echolocate, cats don't) to the cognitive (humans ac-

quire language and monkeys don't). Yet, within mammals at least, it seems quite unlikely at this point that any major species differences exist in the types of ion channels, neurotransmitters, neuronal cell types, tissues, or brain regions that make up a bat, monkey, cat or human brain. At none of these levels of neural organization have neuroscientists discovered a "smoking gun," a neural difference that can account, by itself, for high-level cognitive differences. However, there are both theoretical and empirical grounds for expecting species differences at the cognitive level to derive from differences at the level of the *form* of different (shared) cell types, and the *connections* between these cells. Theoretically, this supposition is based on the fact that genetically-based differences in neuro-computational abilities are implemented by individual cells, expressing their own personal copy of the genome. There is no "mastermind" in neural development, ordering cells to go here and there. Instead, each individual neuron must find its own way, extend its own axons and dendrites based on its own local environmental cues and interactions with other cells. Thus it seems likely that many important genetically-guided differences in neural function will be most directly observed at the cellular level.

For example, the way in which perceptuo-motor connections are made in the developing songbird brain [143,144] is now understood as a time-sensitive interaction between cell adhesion molecules called cadherins. Cadherin expression changes in the post-synaptic motor cells of area RA, facilitating the formation of synapses at the appropriate time. This allows pre-synaptic cells, projecting axons from higher-order associative regions, to form crucial auditory-motor connections of the birdsong control circuit. Thus understanding the implementational architecture of the song system, in genetic and species-specific terms, requires looking at cell-to-cell interactions.

Another nice example of the relevance of cell form comes from recent work comparing FoxP2 gene expression in different species. This gene codes for a transcription factor that turns other genes on and off, and the gene is expressed widely in the nervous system (along with other tissues, such as the lung, where it probably has no cognitive impact). Damage to the FoxP2 gene in humans leads to severe difficulties with the motor control of speech [224,225]. FoxP2 is not "the language gene", but simply one genetic component of a suite of genetic changes that allow our species, but not chimpanzees, to acquire complex learned vocalizations [51]. Although FoxP2 is, in general, a highly-conserved gene, humans have a recently-evolved variant that is shared by all normal humans, but differs from the version found in chimpanzees and other primates [46,47].

Molecular genetic techniques allow both knock-outs of FoxP2 in mouse models [52] and insertion of the human version of the gene into genetically engineered mice [46]. Unsurprisingly, mice expressing the human version of the gene do not suddenly start talking, but they do exhibit a fascinating difference in neuronal structure in a particular population of cells in the basal ganglia. Specifically, medium spiny cells in the striatum exhibit both morphological changes (increased dendrite lengths, and thus presumably increased connectivity and changed integration properties) and physiological changes (increased synaptic plasticity). These specific changes occur in the absence of overall changes in brain size or brain circuitry, and despite the apparent normalcy of these "humanized" mice in a wide range of cognitive tests. The "smoking gun", as far as this particular gene-to-behavior mapping goes, is to be found in the structure and physiology of individual cells, which must somehow "percolate up" to neural circuit and whole brain levels, in ways that remain unclear.

In summary, a long tradition in computational neuroscience and cognitive science more generally has obscured a basic and important fact: the role of the tree-shaped arbor of individual neurons in neural computation. Although computational neuroscience is moving forward with this new insight (e.g., [80]), "cognitive" models have not kept up and for the most part perpetuate the myth of the spherical neuron. Embracing single-cell computation will open a productive route for bridge-building between cognitive, computational, and neuroscientific levels of understanding, and renew attention to understanding how species differences are implemented in neural tissue. Given that the information summarized above represents widely-accepted facts of neuroscience, it does not seem premature for cognitive scientists to begin considering tree-based neuronal networks (rather than networks of spherical neurons) as our basic default model of how biological computation is achieved.

4. Predictive systems theory: what do brains actually compute?

I now turn to what I see as the second major advance in contemporary neuroscience with crucial implications for cognitive science: what I will call "predictive systems theory" but which goes by a variety of names including "predictive coding", "pattern theory", "predictive control", and "Bayesian decision theory" [15,68,156,209,228]. This body of theory is relatively well-formalized, solidly grounded in neuroscience, and encompasses a number of classical

ideas such as the Rescorla–Wagner learning rule and the concept of efference copies. Like the recognition of single-cell computational power, this body of theory is already recognized as both important and empirically well-grounded in neuroscience, but its implications for cognitive science have yet to be widely appreciated (although see [36]). One reason for this slow adoption may be that most of this literature has been couched either in strictly neural terms, or in terms of associative learning theory and not explicitly interpreted in cognitive terms. But as I will show, it would be erroneous to think that predictive principles are in some sense limited to low-level "reflexive" aspects of cognition: prediction-based phenomena are evident throughout the entire cortex, and at all levels of human cognition.

4.1. The predictive insight

The central notion of predictive systems theory is that organisms care less about representing what is actually out there in the world than about how this reality conflicts with their *predictions* about what *should* be there. Put more formally, in terms of the Rescorla–Wagner equation, brains care less about lambda (the data in the world) than about the discrepancy (lambda $- V$), where V represents our current best prediction.

As an intuitive example, consider how the words "tall" and "short" are interpreted. Obviously these words are context dependent in that "tall" means something different for a person than for a building or tree, or even for a man versus a woman. Thus saying "she is a tall woman" connotes that "she is tall, relative to the average woman". In other words, we first take some implicit mean of all (adult female) heights, and then use the *difference* of an exemplar from this mean to determine our notion of "tall". This is perhaps the simplest form of "normalization", where "relative height = actual height − mean height". To improve this measure somewhat, we might perform a two-parameter (Z-score) normalization, where both mean and standard deviation are taken into account. Then we could also encode "very tall", "very short" or "very very tall," in terms of how many standard deviations an exemplar of a tree, person or building are from the mean. Another familiar two-parameter example of predictive coding is the encoding of residuals from a linear regression model. These are just a few of many intuitive examples of predictive coding, where the difference from the value predicted by some model is encoded, rather than the raw values of data.

One virtue of predictive coding of this sort is efficiency via data compression. In the height example above, it will typically require fewer bits to represent the mean and deviations from it (mean = 170 cm, exemplars are $+2$, $+5$, -3) than to store each exemplar in full (172, 175, 167). A widespread practical example is "linear predictive coding", or LPC, a technique used to losslessly compress speech signals [6,124,138]. In LPC, a multi-parametric model of the acoustic signal is first fit to a window of data, and then the residual error from this model is transmitted, along with the model parameters. For the case of human speech, such encoding can lead to major improvements in coding efficiency, with no loss of accuracy, because encoding the error signal typically requires many fewer bits per sample than would the raw data.

4.2. Predictive coding in the nervous system

Based both on such computational considerations, and neuroscientific evidence, it has long been suspected that the nervous system might rely on predictive coding, and encode residual errors relative to models [155,176,196,209]. Couched in terms of sensory perception, we can think of this as a process of unconscious inference [90,148] where higher cortical areas form abstract models, involving imperceptible latent variables (e.g. "there is an animal behind that tree") and then compare the predictions of these models with the actual sensory input. At the neuroanatomical level, Mumford [155] considered the peculiar fact that most regions of cortex have reciprocal connections, not just projecting "upward" to higher regions that appear to process more abstract representations, but also "downward" or "backward" to lower regions that provide its raw data (Fig. 3). Mumford argued that higher regions build an abstract model or "template" of the current scene and then project this template (the prediction) onto lower layers. These lower regions then compute the discrepancy between their actual input and this template back up to higher regions, and an iterative process then attempts to minimize the discrepancy. Mumford hypothesized that this process continues forward from sensory regions to high-level frontal and associative regions of the brain.

A predictive coding framework is able to make sense of several well-known phenomena in neuroscience that are puzzling from other perspectives. For example, many cells in visual cortex can be characterized by sensitivity to some feature (e.g. orientation, color) within a limited receptive field (a small portion of visual space to which they respond). Higher areas then combine multiple features and/or multiple receptive fields. But such cells often

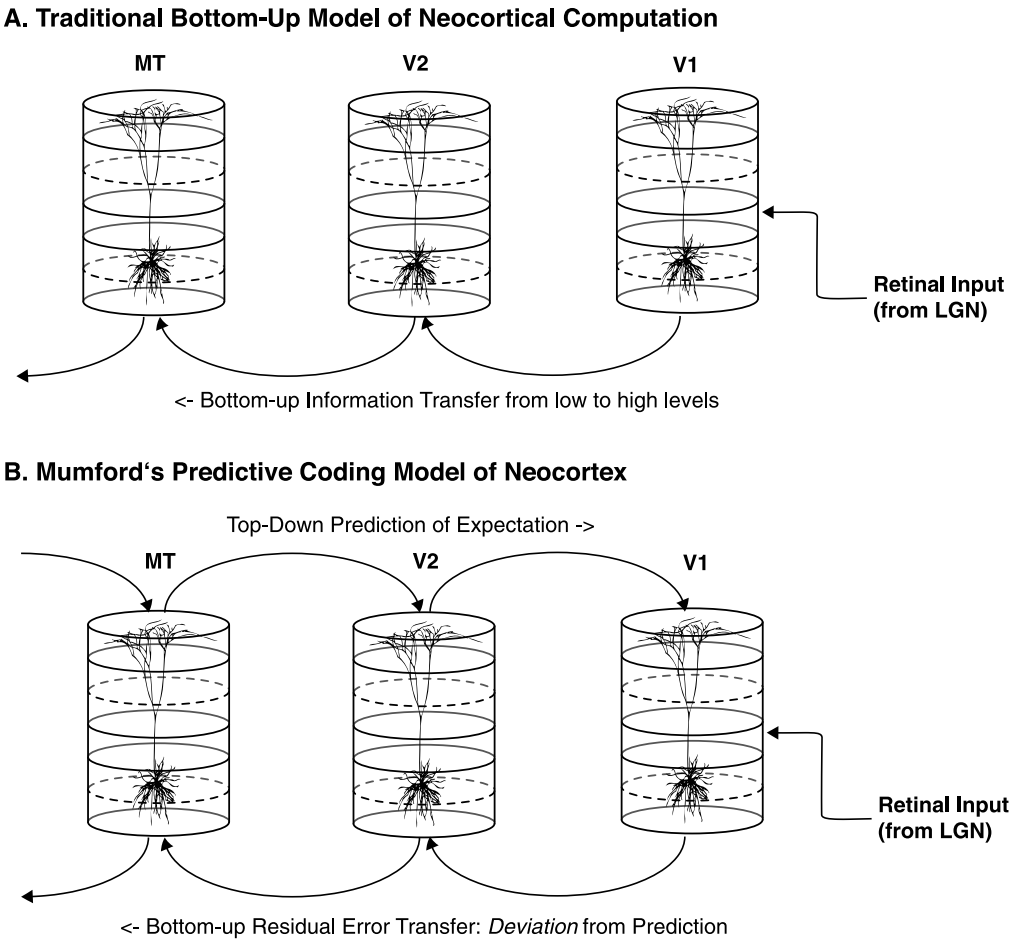

Fig. 3. A schematic illustration of the modern predictive perspective on cortical anatomy, following Mumford's [155] insights, and those of many subsequent theorists. Rather than one way information flow (upper panel), the new model focuses on top-down predictions, and bottom-up signals representing the *deviation* from higher-level predictions (rather than raw sensory information itself).

also exhibit the peculiar phenomenon of "endstopping": an orientation-sensitive cell fires eagerly to a short bar of the correct orientation but decreases its firing to a longer bar extending outside its receptive field [99,100]. Such cells are particularly abundant in layers 2 and 3 of visual cortex, which project mainly to higher cortical areas, and are found both in V1 and higher visual areas (V2, V4 and MT). Rao & Ballard realized that endstopping can be well-explained in predictive terms [176]. If a bar stretches across the receptive fields of multiple cells, higher-level areas can infer its presence, send this "template" back to the lower areas, which then have less residual error and decrease their firing. In contrast, when a bar is short enough to fit within a single cell's receptive field, there is little redundancy with neighboring cells, and a much larger residual from the overall prediction generated by higher areas. Rao & Ballard used these principles to create a hierarchical predictive model whose simulated cells both developed "classical" receptive fields, and showed endstopping.

Analogous predictive principles are now known to describe many other neural systems, including sensorimotor control [68] and the sensing of electromagnetic fields by electric fish [15,89,192]. Electric fish generate weak electrical fields using modified muscle tissue, and use receptors sensitive to tiny modifications of those fields to sense prey and conspecifics. However, changes in the field generated by the fish's own movements are orders of magnitude larger than those generated by the outside world. To compensate, these fish use predictive coding to subtract away the effects that should be generated by their own movements (a form of efference copy). This example of predictive coding provides an interesting evolutionary perspective, because electrical sensing of this kind has evolved at least twice (in mormyriform and gymnotiform fishes, two unrelated clades), but in both cases uses the same essential predictive

principles, implemented in the same brain region (analogous to our cerebellum). Thus, despite the species-specificity of the electrical system itself, the same overarching principles of neural computation independently recur.

4.3. Predictive systems in cognitive neuroscience

Turning now to a more global cognitive level, the well-studied phenomenon of mismatch sensitivity provides a nice example where the predictions occur in time, rather than space [9,72,157]. The brain appears to be exquisitely sensitive to surprises, and these are manifested at the skull surface in changes in the electrical signal measured by electroencephalography (EEG) and/or the magnetic signal measured by magnetoencephalography (MEG). In such studies, it has long been known that surprising auditory stimuli elicit a mismatch negativity (MMN), and that this occurs for a very wide class of repeated auditory stimuli, whenever something unpredictable given the previous events occurs. For example, an MMN can be elicited by a simple series of constant-frequency beeps if beep frequency is changed [215]. However, and more interestingly, if a series of beeps is played where every fifth beep is lower in frequency, the MMN elicited by the low beep quickly disappears, and now re-appears if all five beeps are set to the *same* frequency [230]. These results indicate that the brain is constantly computing residuals from its expectations, at multiple time scales; the MMN is an external indicator of such hierarchical predictive coding. Recent research suggests that this global phenomenon provides a nice bridge to neuronal-level phenomena such as spike-timing dependent plasticity, that can in turn be understood at the level of single-cell phenomena such as long-term potentiation via NMDA channels [229].

Perhaps the most intriguing example of the value of predictive systems theory in cognition comes, as mentioned previously, from the restatement of the Rescorla–Wagner equation in neural terms [194–196]. This research has focused on recordings from dopaminergic cells in the mammalian midbrain, which have very broad axonal arbors encompassing most of the forebrain (both cortex and basal ganglia). Dopamine is an important neurotransmitter for encoding reward, and plays key roles in learning and addiction in humans and animals. The firing rate of these dopaminergic neurons appears, to a good approximation, to directly encode the prediction error term (lambda $- V$) of the Rescorla–Wagner model. Reward translates to firing rate in a signed fashion (decreasing firing rate or increasing it from a baseline rate of 3–5 spikes/sec), enabling the coding of both presence of unpredicted rewards and absence of predicted rewards [195], as required by standard predictive models.

As already discussed earlier, the predictions driving dopaminergic firing are highly cognitive (or as Schultz [195] puts it, "model-based"). Whenever the two conflict, prediction error seems to be based on *subjective* predictions rather than objective ones (cf. [17,142]). For example, dopaminergic neurons react to the *subjective* reward value (discounted for time), rather than the actual amount of food reward. Similarly, when stimuli are presented near sensory thresholds, dopaminergic firing encodes *subjective* presence or absence of stimuli (rather than their actual physical presence or absence). Again, predictive systems theory provides a framework to understand this important marriage of learning theory with neuroscience: Organisms predict what will happen, and they learn most when reality deviates from their predictions. Overall, the brain attempts to model the incoming sensory input by building internal models, and then comparing these models with the current incoming data. When the brain succeeds in perfectly predicting the world, the error is zero, and no learning occurs (and dopamine flows at its normal average rate). When unexpected rewards occur (positive reinforcement), dopamine increases, and the system learns to anticipate variables that might predict that reward in the future. Conversely, when predicted rewards *fail* to occur, dopamine flow decreases, and parameter values that incorrectly predicted success are modified.

I draw two lessons for cognitive science from predictive coding. First, the neural data illustrate the value of quantitative models, like Rescorla–Wagner, for helping to bridge the mind/brain gap. The fact that this model stems from behaviorist learning theory and has traditionally been couched in associationist terms is less important than the fact that it is explicit and makes detailed quantitative predictions that have been tested against experimental data. Such data show that learning is based on organism-internal models of the world, and can be very naturally interpreted in cognitive rather than behaviorist terms. In this case at least, explicit formal models, combined with experimental tests, have achieved what thirty years of qualitative argument after the cognitive revolution (not to mention three hundred years of philosophical argument between rationalists and empiricists) never accomplished: a rapprochement between opposing poles, and one that clearly favors cognitive interpretations. While animal learning theory, based on behavioral data alone, was not enough to achieve this (cf. [180]), its combination with neural data was. This example thus

provides a compelling argument for increased integration of cognitive and neural theory within the explicit cognitive framework of predictive systems theory.

This framework also illustrates the necessity for a clear understanding of the organism's prediction-generating model. The predictive viewpoint stresses that representations are useful insofar as they generate predictions that can be compared with sensory data. From this perspective, cognitive differences among species may be determined mainly by the categories of *model* they are capable of building, rather than how they learn. While Bayesian frameworks and the predictive perspective provide formal specifications of how to use data to update models, neither provides an explicit framework within which models and model-generating capacities can themselves be formulated. In other words, cognitive neuroscience appears to have converged on a useful and neurally-grounded framework within which to consider model-based prediction error, but we still need a framework to make explicit how models themselves are implemented, and how modeling capacities vary from species to species.

To fulfill this need a different body of theory is required: the theory of computation. Despite having been formalized long ago [219] and playing a role in the birth of cognitive science [30,151], its significance for cognitive science, and cognitive biology in particular, has been rather neglected outside of computational linguistics (cf. [60,61]). I will argue that the theory of computation provides the right kind of framework to build and evaluate different types of computational models, and that such a framework is necessary if we want to understand why different species have different cognitive abilities. I suggest that the right way to frame these cognitive differences is in terms of **the classes of models** that different species are able to construct.

5. Understanding species differences

Cell-level computing and predictive coding are aspects of neural computation shared by most animal species. Nonetheless, it is a truism that some species' brains can do things that others cannot: from electroreceptive fish to echolocating bats to language-using humans, different species have different cognitive capacities and propensities well-suited to their ways of life. Species-specificity is as much the norm in animal cognition and behavior as in body form or physiology. Nonetheless, an implicit assumption in much of neuroscience, as well as traditional animal learning theory, is that differences between species are only quantitative: a species may have more or faster learning, or better memory, but the learning rules or memory systems (and thus presumably the neural principles) remain the same. A tension between these views – of qualitative species differences versus quantitative neural differences – pervades the cognitive sciences (particularly where the biology of human cognition is concerned). Unfortunately the debate has not included actual data from a variety of species often enough (cf. [64]) to be well-grounded in biological reality.

We might hope that nervous systems work in essentially the same way across species, if only because we can never hope to understand human brain disease or solve clinical problems without appropriate animal models. The good news, at the neural level, is that many principles of the nervous system, from neurotransmitters to neuronal behavior to developmental wiring principles, are indeed widely shared, and the neural similarities among closely related species (e.g. humans and chimpanzees) seem in most cases to be overwhelming. There is no reason to doubt that the principles reviewed so far (single-neuron computation and predictive Bayesian coding) apply equally to humans and chimpanzees (as well as mice and birds). Nonetheless, chimpanzees do not have language or music, any more than humans have echolocation or electrosensing, so some clear qualitative differences remain. A successful framework for comparative cognition must account for such differences, and attempt to link them to underlying neural differences.

By one account these qualitative differences derive, quantitatively, from differences in brain size. For many years it was suspected that raw brain size, or neuron number, might account for many of the differences in cognitive ability among vertebrates [107,179]. While brain size certainly cannot be ignored as a parameter for understanding the relationship between brain and cognitive capabilities, abundant comparative data show that it cannot provide the complete story. For example, birds typically have far smaller brains than mammals, and yet many perform at a cognitive level meeting or surpassing mammals with much larger brains [45]. Even more surprising, "mini-brains" such as that of the honeybee turn out to be capable of abstract computations, such as the "same/different" distinction, previously thought to require large, complex brains [82]. This suggests, as ethologists have long recognized, that no simple "mass action" principles can explain the details of cognitive capacities, and that details of specific neural circuits are instead crucial.

This provides another excellent reason to put species differences at center stage in cognitive biology. This is an extension of what Krebs has called the "August Krogh Principle": that for most biological problems there is a species in which the problem can be best studied [123]. Finding a closely related pair of species, similar in most ways but differing in some crucial respect, can provide a powerful lens to isolate the mechanistic basis for that difference. By extending Krogh's principle to pairs of species where one possesses and one lacks the trait in question, we can home in on the crucial differences using all the tools of modern neuroscience and genetics. Nice examples of this comparative principle include the use of bird species that are or are not vocal learners to find the neural basis for vocal learning [106,232] or the comparison between monogamous and polygynous voles to discover the neural and genetic basis of male parental care and mate fidelity [101,102,146]. Thus, species differences are not only a reality we must face, but given a good choice of species also provide a powerful and promising avenue to better understand cognition and its mechanistic basis.

How can one approach the question of species differences scientifically? Specifically, how can we determine which of the myriad details that differ from one brain to another actually *make* a cognitive difference? Two classes of answer will be important: motivational and representational differences. **Motivational** species differences concern what is attended to and what sorts of actions come naturally (biting, prehension, flight. . . , cf. [22]). **Representational** differences reflect the type of models an organism is equipped to construct, and how different model components are linked [74]. Such models, from a Bayesian viewpoint, can be used by the organism to bootstrap its probabilistic inferences about the world. Both of these classes of difference will have a strong innate component, but not in any fixed or reflexive sense. Rather, they provide the needed foundation for a bootstrapping process, constraining development in interaction with the environment [58]. Such innate species differences end up resulting, in the adult animal, in complex models that are certainly *not* present at birth, but which nonetheless depend on capacities and proclivities that are part of the species-typical cognitive makeup. Ultimately, we can be hopeful that a mature discipline of computational/cognitive neuroscience will provide mechanistic explanations for both classes of species difference, grounded at the level of neuronal physiology and circuitry.

Although both classes of species difference are important, I will focus here on differences in the modeling capacity of different species. A central goal is to understand the species differences that provide humans with the capacity to acquire music or language, and for these differences I suggest that differences in modeling capacity are of central importance.

6. The theory of computation: a framework for comparative computational neuroscience

Now, presenting the last component of my threefold framework, I argue that the theory of computation initiated by Turing provides an appropriate framework for considering modeling capacity, especially important when considering high-level pattern perception of the sort typifying human music and language. In particular the well-established branch of the theory of computation called formal language theory provides an excellent starting point for experimental investigation of the cognitive abilities of different species. Although I hope deployment of this theory will lead, in time, to a "naturalized" theory of computation in living tissue, we are far from such a theory at present.

6.1. The theory of computation and formal language theory

There are two main components of the modern theory of computation: formal language theory (FLT) and the theory of algorithms (TOA). Focusing on FLT, I provide a short non-technical overview below, since the discipline is well-covered in widely available computer science textbooks [78,97,130,131] and shorter recent reviews [60,105]. Accessible popular treatments are also available [16,42].

FLT is essentially the study of infinite sets generated by finite means, a characterization that makes its significance for music or language clear. The field had its start in pure mathematics in the early 20th century, when mathematicians sought to understand how the (presumably infinite) set of true theorems and correct proofs, all building on a finite set of mathematical rules, could be generated. A crowning achievement of this endeavor was Turing's model of computation, now called a Turing Machine [219]. Although there were at the time multiple competing proposed models for computation including Emil Post's rewrite systems [173] or Alonzo Church's lambda calculus [34], it rapidly became clear that all of them can generate the same infinite set of propositions/proofs, and thus were all formally equivalent. Although each variant is still used today, it is now common to use the Turing machine as a

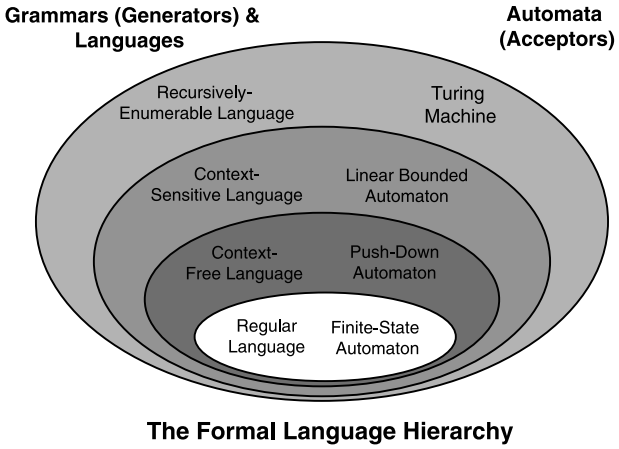

Fig. 4. The traditional formal language hierarchy is a nested inclusion hierarchy which allies classes of grammars (which are finite rule systems) and languages (the potentially infinite string-sets they generate) with the computational machinery minimally required to calculate that class.

stand-in for all of these, and indeed all imaginable, formal computational models. "Turing equivalence" thus defines the outer reach of anything that can be clearly specified as an algorithm.

In the meantime another body of theory was developing, adopting a more practical or statistical perspective, that included simpler models such as Markov processes and regular grammars [114,115]. Unlike the all-powerful systems of Turing, such models were explicitly limited in power and complexity, both to allow better understanding and aid practical applications. These again constitute a large class of computational systems, sometimes called finite-state systems [32], which again can generate infinite sets of strings using a finite set of rules, but in this case which have clear limitations. For instance, the simple string pattern A^nB^n, in which the number of As is precisely matched by the number of Bs, can easily be shown to be beyond the capability of any finite-state system [152], although any such string could be identified, given adequate time and patience, by a human crossing off alternate As and Bs, one by one (or equivalently by a person counting the number of As, and the number of Bs, and comparing the final counts). Other more interesting limitations are that finite-state systems cannot recognize mirror symmetry in an arbitrary string (e.g. "abc|cba"), or even simple repetition of an arbitrary string ("abca|abca"). Thus, despite their considerable power, finite-state systems do not appear to be adequate to model human mental capacities *in toto*.

6.2. *Supra-regular grammars*

These two levels of computational power (finite-state and Turing-complete) define the innermost and outermost levels of the modern formal language hierarchy (Fig. 4). But it became clear in the 1950s that these alone are too coarse for categorizing the many various algorithms that began to arise as computers became more powerful, or the problems raised by considering human languages as infinite sets of sentences generated by finite means. This led Noam Chomsky and his colleagues to define two intermediate levels of power, captured by so-called "context sensitive" and "context free" grammars. The latter, abbreviated CFGs, are particularly important from a modern comparative perspective because (as we will see below) they appear to embody a type of tree-oriented computation that humans find very natural, but that other species find challenging or impossible [60,150].

Context free grammars are classically defined, following [173], in terms of "rewrite rules" which allow one string to be rewritten as another string, given certain restrictions. Different classes of grammars are defined by allowing some types of conditions and not others. For example, a context-free grammar can include rules such as

$$A \rightarrow A\ B\ C$$

which says to rewrite the string 'A' as the string 'A B C'. However, a rule such as

$$zA \rightarrow A\ B\ C$$

which indicates that A can be so rewritten *only* when preceded by a z is called "context sensitive", and such rules are forbidden in CFGs. This restriction provides a powerful bonus: it means that it is easy to build fast, efficient

parsers for CFGs, in contrast to unrestricted context-sensitive or higher systems [97]. Thus, all modern computer languages are based on CFGs, which have become an indispensable practical tool in computer science [78,131,190]. Context-sensitive grammars have fewer practical applications.

Adding these two new classes of grammars, intermediate between the finite-state and Turing-complete levels, leads to the "standard" formal language hierarchy presented in textbooks (often called the Chomsky Hierarchy – though it would be equally correct to call it the Turing/Chomsky/Kleene hierarchy). However, this was not the end of the story: at each level of this hierarchy finer distinctions have now been made, including a finer subdivision of the finite-state zone into the so-called "subregular hierarchy" [105]. It also turns out that, although most components of most human languages can be captured relying solely on finite-state and context-free representations, multiple languages exhibit features that go slightly beyond the power of these systems (in particular, so-called crossing dependencies in Dutch or Swiss German). This led to the codification of an important new supra-regular level: the "mildly context sensitive" (MCS) languages [201,210]. In an interesting reprise of the original convergence of the Turing/Church/Post approaches to computation, multiple contemporary formalisms for writing natural language grammars turn out again to converge at this new level, and are thus formally (weakly) equivalent [111,227]. The formal language hierarchy is thus not a fixed edifice, but a work in progress: an extendable roadmap for understanding what types of rules and computations are available to systems with particular capacities and limitations.

A crucial virtue of the theory of computation is that it defines formal equivalence classes of several sorts. First, as already mentioned, it provides a well-codified sense in which two approaches can be shown to be "formally equivalent". In the cognitive sciences, we need a shared framework for deciding if two superficially different modeling frameworks are simply notational variants of one another. Once we know this, the choice between the frameworks becomes a matter of taste and convenience rather than vehement debate. Examples of such notational variants include modern programming languages, Turing machines, and Post rewrite systems (all are Turing equivalent) or multiple modern grammatical formalisms (such as minimalism, tree-adjoining grammars, and combinatory categorial grammars) which all predict the same MCS linguistic extensions [210,227].

A second important type of equivalence occurs between rule systems and processing algorithms – between abstract descriptions of rules and the machines ("automata") that can compute them [97,131,153]. If we can show that a particular problem can be described with a regular grammar, Kleene's theorem tells us we can solve that problem using a finite-state automaton [115]. Thus, once we can define the problem, we can say what type of equipment we need to solve it. Correspondingly, if we can model a given machine at the formal automaton level, we automatically know about its computational capacities and limitations.

6.3. *The supra-regular hypothesis for human music and language*

A founding insight of both cognitive science and modern linguistics is that all human languages require supra-regular computational resources (resources above the finite-state level) [29,30]. This means that although a finite-state automaton can solve many useful problems (e.g. learn a lexicon, recognize word strings, etc.) there is a substantial class of problems that it cannot solve. These include all aspects of language in which flexible, extendable trees are needed as data structures, or where tree-identification and processing are core computational problems. For systems which rely strongly upon flexible, nested hierarchical structure, as do language and music, such tree-based processing is indispensable. Thus, most current computational linguistic models of language take the existence of both tree structures, and context-free processing resources, for granted (e.g., [110,164]). Similarly, modern computer language compilers typically involve two steps: a finite-state "tokenizer" (or word recognition system) and a context-free "parser" which recognizes syntactic structures built up from these tokens [2,3].

Evaluating the results of the earliest experiments exploring human artificial grammar learning, psychologist George Miller reached the conclusion that human participants appear to have a bias to employ context-free solutions *even when these are not required by the data*. In Miller's words:

> "constituent structure languages are more natural, easier to cope with, than regular languages… The hierarchical structure of strings generated by constituent-structure grammars is characteristic of much other behavior that is sequentially organized; it seems plausible that it would be easier for people than would the left-to-right organization characteristic of strings generated by regular grammars"
>
> [150, p. 140]

where "constituent structure languages" refers to supra-regular (context-free or above) grammars. Thus, updating to use modern terminology (cf. [61]) we can restate George Miller's conclusion as the **Supra-Regular Hypothesis** for human cognition:

> When presented with sets of strings, humans have a both a capacity and proclivity to infer hierarchical structures wherever possible

Because the term "hierarchical" can be interpreted in many ways, we need some definitions in order to make this hypothesis precise.

6.4. "Hierarchy", "trees" and context-free grammars

Here are two definitions to clarify more precisely the interpretations employed here:

Hierarchical Structure: A structure whose graph takes the form of a rooted tree

Rooted Tree: An acyclic, fully-connected graph with a designated root node

Although regular grammars can construct trees, they are of a depauperate sort [129]: either exclusively right- or left-branching. In other word, all recursive rules in a regular grammar must have the same form e.g.:

$$S \rightarrow a\ S$$

which allows them to generate "tail-recursive" trees that branch indefinitely off to the right. In contrast, a context-free grammar also allows rules of the form $S \rightarrow S\ a$, or

$$S \rightarrow a\ S\ b$$

It is important to note that not *all* rules of a CFG need to have these forms: a single one is enough to render the entire grammar supra-regular. The crucial advantage allowed by such rules is that they can generate trees with *any arbitrary structure*. Right-, left-, or both branches are allowed, and CFGs by no means entail strictly center-embedded structures. Thus, any syntactic system allowing a rich variety of tree structures is most compactly described in supra-regular terms. All natural languages, and most modern computer languages (C, Java, Python, ...), fall into this category.

Although our understanding of the computational nature of human music is not as advanced as that of language, most theorists agree that music, like language, possesses rich hierarchical structure at multiple levels (e.g. melody, rhythm and harmony) and that this is a central computational similarity between music and language [96,119,127,128,133,141,161,183]. Like language, music makes "infinite use of finite means" by combining a limited set of atomic notes into riffs, measures, phrases, and movements of greater and greater scope. These arbitrary hierarchical structures require supra-regular grammars to be concisely notated [183,212]. Miller's hypothesis can therefore be tentatively generalized to include music. The supra-regular hypothesis thus has a multi-domain scope, including at least language and music.

In summary, Miller's supra-regular hypothesis holds that *humans* confronted with a set of strings first attempt to find some general rules by which hierarchical structures can be inferred for those strings. The theory of computation allows us to conclude both that such rules require a modeling system with supra-regular capacities (i.e., a context-free grammar or better), and that such a system requires computing machinery with powers beyond that of any finite-state machine (e.g., requiring a "push-down automaton" or better [78,97]). But what of other species? Do non-human animals also infer tree structures when confronted with strings?

6.5. Supra-regularity and animal cognition

In 1999, as a post-doc at MIT, it occurred to me that formal language theory could be employed as a framework for animal cognitive experiments designed to answer precisely this question. At that time, ongoing experiments with human infants using the paradigm of artificial grammar learning (AGL) were producing exciting insights into the pattern-recognition capacities of infants [137,188]. Artificial grammar learning, in its classic form [177], involves

generating a set of strings using some set of rules (the "grammar"), and exposing participants to members of this set. The subsequent "test phase" involves presenting novel grammatical strings and ungrammatical violations, to determine what rules the participants acquired (if any). For infants and/or animals, this paradigm can be combined with habituation/discrimination techniques to provide a powerful empirical platform for pattern learning studies. Habituation techniques make use of the simple fact that organisms repeatedly exposed to similar stimuli habituate to them (itself a reflection of the predictive systems approach to the brain/mind). When a novel stimulus differing from predecessors is then presented, the organism will "dis-habituate" if it recognizes the novelty, and show a renewed response. Since dishabituation requires that the organism both infer some commonality in the previous stimulus set, and recognize that this commonality is lacking in the novel stimulus, it provides an excellent way to probe pattern learning and discrimination in a nonverbal organism, with no need for extensive training [28,43].

Despite a very large literature using AGL, there had been very little attention to the regular/supra-regular distinction since Miller's initial "Grammarama" experiments [149,150]. Indeed the vast majority of human artificial grammar research used roughly the same grammar as that introduced by Arthur Reber in the first AGL research [177] – a moderately complex finite-state grammar. I thus began considering performing AGL experiments with artificial languages going beyond finite-state. Although there are many supra-regular candidates, I first thought of the context-free grammar A^nB^n as a simple grammar which generates such strings as {AA BB, AAA BBB, etc.}. The recognition of this grammar is beyond the capacities of finite-state automata, because such machines have no way to keep track of an arbitrary number of past As in order to match them with the same number of Bs. I first tested a few human participants to ensure that they could master such a grammar, which they did without apparent effort (for both tone and syllable strings). I then teamed up with Marc Hauser at Harvard, who ran a lab already testing cognition in cotton-top tamarin monkeys (*Saguinus oedipus*), including AGL experiments. As a control we used the regular grammar $(AB)^n$, which generates strings like {AB, ABAB, ABABAB, ...}. When exposed to this regular language, the monkeys were able to recognize the pattern, as evidenced by more frequent looks to novel stimuli violating this pattern. In contrast, despite repeated testing with several different string types (tones, syllables), these monkeys failed to recognize the supra-regular A^nB^n grammar [62].

The publication of this study launched a wave of further experimentation combining AGL and FLT, in both humans and multiple animal species [60,214]. In an unfortunate terminological confusion, the issue of regularity was confused with that of "recursion", such that many researchers have attempted to use success on the A^nB^n grammar as evidence for recursive processing in animals [1,77,136,181]. Similarly, some researchers have confused the capacity for context-free grammars as implying pure center-embedded structures [166,181], rather than the flexible and general class of structures that CFGs are capable of. Center-embedded sentences nest phrases in the midst of other phrases, and are well-known to be very challenging to parse beyond two levels of embedding. For example, in the sentence "White is the color that the painter that Fred hired prefers" involves embedding the phrase "that Fred hired" in the center of the phrase "the color that the painter prefers". Because this process cannot be extended indefinitely without a loss of comprehension ("White is the color that the painter that the neighbor who Fred hates hired prefers" is already borderline), center-embedding has long been considered to be a borderline phenomenon in psycholinguistics, perhaps grammatical in theory but unparsable in practice [7,152]. But nothing limits context-free grammars to such center-embedded structures: a flexible mixture of right- and left-branching and center-embedding is possible. Nor does A^nB^n need to be center-embedded (with links between particular As and Bs) to be context-free (see below, and cf. [60]).

Such misunderstandings and confusions are perhaps to be expected in the early days of a new paradigm, but it is unfortunate that they have obscured what I take to be the main conclusion of this body of research: that those non-human species that have been tested so far do not seem to have the same propensity as human subjects to perceive strings as hierarchically structured, and as a result find the simple but supra-regular A^nB^n language difficult, or impossible, to identify. This appears to be true regardless of the sensory domain, and occurs whether the strings are made of human speech syllables, conspecific vocalizations, or abstract visual tiles [62,213,223]. In contrast, in many different experiments with a variety of stimulus types, humans easily and intuitively master A^nB^n and other supra-regular grammars [8,41,62,71,95,166,220].

6.6. The Dendrophilia Hypothesis

Although many species and many grammars remain to be tested, the current results thus provide tentative grounds to consider an extension of Miller's supra-regular hypothesis (which concerned humans alone) to one with a broader

comparative scope. This hypothesis seeks to characterize the cognitive difference between humans and (some) other species in terms of the class of models that we can build to make sense of stimuli. In particular, I suggest that humans have a species-typical, multi-domain inclination to infer tree structures as our underlying model for string sets, but that this is *not* the case for most animal species.

Because the core notion here is that "humans love trees" I call this broader hypothesis
the **Dendrophilia Hypothesis**:

> "Humans have a multi-domain capacity and proclivity to infer tree structures from strings, to a degree that is difficult or impossible for most non-human animal species"

This hypothesis incorporates the core observation in Miller's supra-regular hypothesis, but explicitly limits the hypothesis phylogenetically, suggesting that our dendrophilia is unusual from a comparative perspective. The dendrophilia hypothesis explicitly extends human supra-regular proclivities to non-linguistic domains such as music (and, more tenuously, visual pattern perception, cf. [231]). To the extent that this hypothesis is correct, it provides a unified *computational* account of the difference that allows our species, and not others, to acquire language: that we infer trees over linguistic or musical strings, and that this allows us to compute probabilities and infer higher-order rules that would be essentially invisible to a species that does not do so.

Summarizing, with the dendrophilia hypothesis I attempt to isolate and characterize an important cognitive difference between our species and others which, I argue, helps explain why a human infant easily acquires language but a cat, dog or monkey raised in the same household does not. Although some animals can acquire large vocabularies, e.g. several hundred object labels in dogs and chimpanzees [84,113,167,191], these species appear to be limited in their capacity to infer larger-scale hierarchical syntactic structures over these word strings (cf. [191]). In the context of the (broadly-shared) Bayesian framework discussed earlier, this would mean that probability distributions over trees simply are not computed by dogs or chimpanzees, not because they can't compute probabilities, but because they don't infer the trees in the first place. Given the centrality of tree-centered probability distributions in modern computational linguistics, this would constitute a fatal limitation in acquiring a human language. This limitation presumably results from differences in brain structure between humans and animals that we will consider in the final section.

6.7. Employing formal language theory in behavioral experiments

Before turning to some of the implications of the previous discussion, it is important to forestall several areas of confusion which have caused unnecessary and unproductive debate in the past. The first is the mistaken notion that the research programme on animal AGL described above concerns "recursion". Because recursivity is at the heart of any system that generates infinite sets with finite means, *every* level of the formal language hierarchy by definition permits recursive rules. In a finite-state grammar such recursive rules are limited to the form $S \rightarrow S\ a$ (so-called "tail recursion"), and the structures they build are uniformly right- or left-branching trees. But they are defined recursively nonetheless. In contrast, the A^nB^n grammar can be simply defined by defining two counters for the number of As and Bs, and comparing these numbers when the end of the string is reached. Although this approach is provably supra-regular, and cannot be captured by a finite-state automaton, it does not require what would typically be defined as a recursive definition. The contrast between finite-state and supra-regular grammars is thus a poor test of recursivity, and was never proposed as such in the original Fitch & Hauser [62] paper. In short, "recursion" and "supra-regularity" are independent axes along which a computational system can be classified, and should not be confused (cf. [56,60]).

Another issue concerns a misguided criticism that A^nB^n can be solved trivially by "just counting". While, as just observed, this grammar can indeed be solved via counting, this is by no means a trivial fact because, as proven by Minsky in 1967, an automaton possessing two integer counters plus some simple increment and comparison operators is in fact Turing equivalent [153, Chp. 14]. In the theory of computation, two counters and a comparison operator go a long way, and such resources are far from trivial, since a few more operators generate a Turing-equivalent computational system.

A related confusion concerns "center embedding". The central difference between a context-free and regular grammar concerns the form of recursive rules allowed, which indeed allows CFGs to build, among other things, center-embedded structures. However, CFGs can generate trees using any combination of right-branching, left-branching

and center-embedded structures. Thus what makes these grammars important is their generality: their capacity to generate *arbitrary* tree structures. There was a flurry of debate about whether human subjects automatically infer center-embedded structures for A^nB^n strings, after [166] found that their human subjects did not. First, multiple studies now show that humans are perfectly capable of inferring and extending such center-embedded structures, when provided with the proper training and exemplars [8,220]. Second, for reasons just explained, a failure to acquire such structures does not imply that these subjects failed to learn A^nB^n! Finally, in addition to the "count and compare" and center-embedding solutions for the A^nB^n language, it is equally legitimate to posit "crossed" dependencies, of a sort that can also be inferred by human participants [220]. In short, it is an error to think that the supra-regularity of the A^nB^n language, which concerns a set of strings, hangs on any particular structural interpretation of those strings. Center embedding is indeed an interesting phenomenon, but researchers interested in investigating it should design experiments using mirror or palindrome grammars, where center-embedded structure is explicitly required, rather than continuing to investigate A^nB^n.

The final, and more interesting, empirical issue concerns the type of evidence required to conclude that a learner has in fact acquired a supra-regular grammar. Showing this involves more than simply showing above-chance recognition of strings generated by a given grammar, because there are often regular grammars (or combinations thereof) that can successfully accept most A^nB^n strings and reject many others. For example, in a comparison limited to A^nB^n and $(AB)^n$ grammars, several simple regular patterns can distinguish between these strings (for example, simply accepting the bigrams 'AA' and 'BB' will allow AABB or AAABBB strings to be accepted, while rejecting ABAB or ABABAB strings). Thus, it is necessary to design a careful set of probe trials that are beyond reach of such simple strategies. In this case, the crucial control condition is composed of "mismatched" A^*B^* strings where the number of As and Bs is unequal [60]. Because such patterns cannot be rejected by most regular strategies, successful rejection of such patterns by a subject is required before the conclusion of supra-regularity is warranted [77,213].

A related issue concerns generalization over n, which is again required to infer a supra-regular strategy. While we don't expect any organism (human or otherwise) that has acquired the A^nB^n grammar to accept string with arbitrarily large ns (e.g. 100 As and 100 Bs), we do expect some generalization over whatever ns have been observed. That is, if animals are trained with $n = 2$ or 3 (e.g. A^2B^2 and A^3B^3 strings) they would need to generalize over n, accepting patterns of the form A^4B^4, before we could conclude that they have learned A^nB^n. This is because the union of two regular grammars – $A^2B^2 \cup A^3B^3$ – is still a regular grammar, and there is no reason to expect a subject employing it to generalize it to accept A^4B^4 strings [62,77,213]. Thus, an empirical demonstration of supra-regularity for such grammars involves both generalization to new stimulus lengths (generalization over n in the case of A^nB^n) and rejection of carefully chosen "foil" stimuli (for that grammar, strings with mismatched number of As and Bs). Such minimal requirements have not been met by some recent experiments [166,181], and this renders unsubstantiated the claim that "baboons parse supra-regular grammars" (much less that made by the authors that "baboons have recursion"). For critiques of other recent claims of animal supra-regularity see [14,214,223].

6.8. Summary

I conclude by recapping the overall point of this section: formal language theory provides a rigorous, explicit framework (along with a concise notation and a large body of theorems and practical experience) within which core questions concerning the nature of neural computation can be framed and more importantly *tested* in multiple species. The current data support Miller's supra-regular hypothesis concerning humans' ability (and propensity) to infer hierarchical structures over strings, and are consistent with my dendrophilia hypothesis: that humans but not most other species habitually and easily infer trees from string sets. But although the recent debate has been focused on the finite-state/supra-regular distinction, this should not blind us to the fact that there are many other regions of formal language space worth exploring. Current data suggest that the sub-regular language hierarchy may prove more fruitful in the long run, and may be particularly relevant to the domains of phonology and for many aspects of music.

Why should the capacity to represent and manipulate trees represent such an important difference between species? The answer has to do with representing hierarchical systems in an efficient and extendable way. A hierarchical representation is "chunked" into subparts, with the labels for the subparts corresponding to higher-level nodes in the tree. Such a representation allows these higher-level abstractions to be repeated, reorganized and generalized, allowing greater flexibility than a simple sequential list of elements [206]. Complex motor planning (e.g. for sophisticated tool making), language and music all involve tree representations, because they all rely on hidden but important

intermediate-level organization that must be inferred if the visible sequential surface (the terminal "leaves" of the tree) is to be correctly parsed or executed. Although sequences can of course be memorized into a simple list, parsing them into tree structures with hidden higher-level nodes can provide a more compact "description language" for a large set of sequences than does a simple list. Tree-based representations also allow flexible and creative extension of the system by rearranging sub-trees. A system that has inferred the higher order hidden structure {subject–verb–object}, after learning and parsing "Jim chases John," can easily generate "John chases Jim" as well, even if that sentence has never been heard. Thus, I argue that the capacity to represent and manipulate trees is a core human ability, shared across multiple cognitive domains, and valuable because it provides a "description language" that is both highly compressible and broadly extensible.

7. Synthesis: a neuro-computational framework for a cognitive biology of music and language

I conclude with an illustration, putting the proposed neuro-computational framework to a more specific use, focused on the biology of music and language. The overarching question is one that has occupied a central place in cognitive science since its origin:

"What are the biological pre-requisites for language acquisition?"

This has been a recurrent (and polarizing) question for many decades. The ultra-behaviorist answer [208] was essentially "nothing more than standard conditioning, and the right environment", while the answer from generative linguists has long been "a suite of special purpose systems, and a good dose of innate knowledge" [31]. But underlying these apparently irreconcilable viewpoints are large areas of agreement: everyone recognizes that a kitten, raised in a human home, will not learn language but that a normal child will. There can be little doubt that this difference has a strong genetic component, which results from differences between the nervous system of kittens and babies. The key questions concern the computational nature of the prerequisites for language acquisition, how they support learning (at least of vocabulary), their neural implementation, and which pre-requisites are shared with other animals. My colleagues and I have suggested elsewhere that *most* of the cognitive pre-requisites of language are in fact shared with other species, in some cases many other species [54,55,88], although this claim remains controversial [104,170]. What is not particularly controversial is the proposition that, despite massive similarities in brain function between different species, some features of human brains enable us and not other species to acquire language. What these features are is clearly an interesting question to which we would like to know the answer.

In this last section, I will explore some approaches to tackling this question, using the framework advanced above, and proposing several specific testable hypotheses. To forestall confusion and misunderstanding, note the important distinction between the framework detailed in the preceding pages, and the more speculative and tentative proposals made below. The value of the framework should be decided on its own terms, and not tied too hastily to the fate of the suggestions that follow. These suggestions may well be wrong, but the way to find out is to employ a rigorous, computational framework – my main focus above.

7.1. Language, tree-building and the fronto-sensory scratchpad

The framework proposed here puts a central focus on tree-processing operations as a core necessity for music and language, and posits supra-regular processing power as the corresponding computational abstraction. It is natural to ask how such a system is implemented in neural hardware, within a predictive systems framework [70,174]. While the answer to this question will remain tentative until a viable *animal* model for tree-processing is identified and studied, the best current candidate based on human research involves the bilateral peri-Sylvian network connecting the inferior frontal gyrus (IFG, comprising Broca's area and its neighbors), with sensory and association regions in the temporal and parietal lobes. Unlike the "standard" cortico–cortical interconnections highlighted in the Mumford model between adjacent cortical regions, these represent long-distance connections connecting primitively pre-motor areas with sensory regions. These dorsal pathway connections (including the arcuate fasiculus and superior longitudinal fasiculus [69]) have undergone considerable expansion and modification in recent human evolution, as evidenced by differences compared to both chimpanzees and macaques [182].

The first indication that something may be odd about the "motor" regions of the IFG came from the now well-documented observation that Broca's aphasics have difficulty not just in producing but in *perceiving* syntax [25], especially when semantic information is not available to disambiguate structure. With the rise of brain imaging it became clear that IFG activation is a typical aspect of structural processing in perception of both music [120,121,135] and language [20,70,87,94]. This finding is well-established, but we must nonetheless ask why a "pre-motor" region should play such a central role in an essentially perceptual process (cf. [202]).

One possibility is that the IFG serves as a kind of "abstract scratchpad" for the sensory predictive engine ensconced in the occipital and temporal cortices, allowing these to "offload" partial results computed during serial processing of hierarchical structures (cf. [65,174]). Thus, reverberations in the fronto-sensory feedback loop would play the role of the stack in the pushdown automaton implementing a context-free grammar. In addition to the standard hierarchical predictive engine, locally implemented in all sensory cortices, the posterior regions connected to the IFG would thus have an additional storage mechanism into which intermediate results (and in particular unfinished structural computations) could be placed for later retrieval. Consistent with this idea, IFG areas are increasingly activated as the size of the hierarchical structures processed increases, and this is true even with "Jabberwocky" sentences made up of nonsense content words, but incorporating normal function words (indicating grammatical structure) [159].

This IFG-as-stack model suggests that structural processing does not occur in Broca's region *per se*, but rather that a crucial aspect of this processing (the stack-based storage of intermediate results) is accomplished via a network of which Broca's area is an important component. Note that the activation of the *right* inferior frontal region in musical syntax tasks is perfectly consistent with this hypothesis, and that this language/music parallel strengthens the notion that the fronto-sensory loop is centrally involved in structural, rather than semantic, processing (which is the focus of a ventral temporal lobe network). I have suggested elsewhere that the evolutionary origins of this fronto-sensory loop may be found in selection for complex vocal learning, which requires intimate informational exchange between auditory (temporal) and motor/pre-motor regions [57]. By this hypothesis, sensorimotor circuits that initially evolved for acoustically driven motor control were "exapted" later in our evolutionary history for use as an abstract fronto-sensory scratchpad, and this neural analog of a pushdown stack then became available for domain-general structural computation of the sort typical of music and language.

While Broca's area *per se* (BA 44/45) seems to be preferentially involved in phrasal syntax, surrounding areas play roles in both lower-level structure (phonology) and higher-level meaning (semantics). It thus seems prudent to consider the entire IFG, and not just Broca's region, as the fronto-sensory scratchpad (cf. [86,87]). Nonetheless, it is quite interesting to note that BA 44 and 45 specifically have undergone a spectacular enlargement in recent human evolution. Detailed cytoarchitectonic analysis of these areas in chimpanzees and humans indicates that BA 44 and 45 are present, in comparable locations, in chimpanzees (contra some earlier suggestions, e.g. [160]). However, these regions are among the most greatly expanded cortical areas yet identified in humans: left area 44 is 6.6 times larger in humans than in chimpanzees and left area 45 is 6.0 times larger [193]. This is disproportionate compared to the overall increase in brain size: the average human brain is only roughly 3.5 times larger than that of a typical chimpanzee, and V1 is only 1.8 times larger in humans than in chimpanzees.

In summary, by this hypothesis, a direct link is made between the computational need for an augmented memory system (stack, queue or equivalent) in a supra-regular automaton, and Broca's area, a region that has undergone considerable expansion and rearrangement in recent human evolution. This additional memory system is implemented via a fronto-sensory scratchpad comprising the inferior frontal gyrus (centered on Broca's area) and posterior sensory regions, and by the enlarged long-distance fiber tracts connecting them. This hypothesis is consistent with "standard" models of the peri-Sylvian language regions [79,94], but goes beyond them in drawing a more specific link between such neural circuitry and computational theory.

7.2. Identifying the neuro-computational changes underwriting structure processing

The hypothesis above specifies a cytoarchitectonically-defined region as having apparent importance in a well-defined computation (stack-based tree building), but does not say *why* this region of cortex should play an important role. While a partial answer may simply be "because it has the right connections to other brain regions", it seems likely that the differences run deeper than that. For example, the IFG continues to play a key role in signed language, despite a completely different signaling modality, so it seems unlikely that the IFG circuit's importance derives solely from its strong connections to (auditory) temporal cortex. I suspect that another part of the answer derives directly from

the original pre-motor functions of the IFG, which must influence the type of information flow within the multiple cortical regions it contains.

Another possibility, not mutually exclusive, is that relatively subtle changes in cellular morphology within the IFG lead to a phase transition in the types of computations the whole circuit can compute. This brings in the third, neuronal computing, aspect of my framework. Consider again the changes in spiny cell morphology that result from inserting the human form of the FoxP2 gene into genetically-engineered mice [46,226]. These cells, within the striatum, have significantly longer dendrites, and thus more extensive dendritic arbors than those in wild type mice. If something similar happened in cortical pyramidal cells, it could serve to increase the receptive field size (spatial and/or temporal) of these neurons, and of whole cortical columns. Such subtle morphological changes, and/or changes in the expression pattern of particular ion channels within that dendritic arbor, could change the time constants of neurons, with the effect of enlarging their temporal receptive fields. Such changes would allow frontal neurons to process longer time chunks (and thus perceive longer-distance relationships) than a system with a narrower receptive field. Such an increase might lead in turn to an increase in the *depth* of abstract trees that could be represented (e.g. the number of levels of branching), and here even a single additional layer could add very considerable additional processing and expressive power.

While surely speculative, this hypothesis is explicit, and consistent with what little we know about how changes in DNA could effect neural computation via changes in cellular computation. It is also consistent with the broader principles of neural computation discussed above. In particular, this hypothesis highlights the potential importance of single-neuron computational properties in generating higher-order differences in the computations performed by larger brain regions. I strongly suspect that such explanatory cell-to-circuit principles are an important part of the bridge-building between neuroscience and cognition, even if the details of the hypothesis above turn out to be incorrect.

7.3. Evolutionary origins: where do the constraints on "Universal Grammar" come from?

Finally, consider another highly contentious issue in the cognitive sciences: the degree to which humans can be said to have "innate knowledge" crucial for language acquisition. Generative linguists have termed this hypothetical knowledge-base "Universal Grammar" [31,103,158]. It is important to note that Chomsky's Universal Grammar was never intended to denote properties shared by all of the world's languages (contra [48,49]). Rather, the term was introduced to denote the deeper cognitive principles underlying the acquisition of any language, presumably shared by all normal humans (cf. [59]). Nonetheless, both "innate knowledge" and "Universal Grammar" have proven exceedingly controversial, so much so that it is difficult to find calm discussion of the various issues raised by these concepts.

I personally find the terms "innate knowledge" and "Universal Grammar" somewhat misleading, with unfortunate implications that often obscure the crucial issues. Thus I will not defend the terms themselves, but instead concentrate on the important concepts underlying this long-running debate. Two key issues need to be clearly distinguished:

1) Constraints: What kinds of **biological constraints** are present in the child learning language, such that children so reliably converge on the correct meanings and structures of their language(s), and apparently avoid ever considering a huge number of logically possible alternative possibilities?
2) Species Specificity: What **novel capacities and proclivities** (e.g. novel representational formats) are available to the human mind, and not found in related species, that support the language acquisition process?

It is important to note that question one, concerning constraints, implies nothing about the constraints being specific to language, or specific to humans. The answer to this question could involve many inherited, reliably-developing ("innate") constraints which are *shared with other species*, and/or important in non-linguistic cognitive domains. In confronting the massive task of language acquisition, the child needs all the help it can get, and there are certainly many constraints available from ancient neural circuitry (e.g. for vision, audition, motor control, and social cognition) that could help disambiguate reference, avoid unrealistic or unnatural hypotheses, and generally speed the language-learner's progress towards parsing and understanding the language(s) of its community. They needn't be human- or language-specific. Thus, many important components of our biological endowment for language learning may be shared with other animals [63,88]. This is not a new idea: Chomsky [31] proposed "that proper names... must designate objects meeting a condition of spatio-temporal contiguity" or that "color words of any language must subdivide

the color spectrum into continuous segments" are plausible examples of formal universals, unrestricted to syntax or language. Indeed, the color example is clearly tightly bound to the sensory world of vision.

To make this possibility more explicit, I suggest that the human child brings a very significant biologically given set of constraints to the language acquisition problem, constraints playing a central role in reducing the lexical, semantic and syntactic hypothesis space the child implicitly explores. By this proposal most if not all of these constraints are widely shared with other species. Indeed these innate constraints have evolved, like the vertebrate brain itself, over more than 500 million years to enable organisms to rapidly and effectively convert perceptual stimuli into useful predictive models about the environment. These constraints almost certainly represent a complex mosaic of many factors, all of them reliably developing in mammalian brains, in general. Thus the answer to question one, above, involves a large, complex and ancient set of Bayesian priors (visual, sensory, motor) that constrain inference in any mammalian brain, and are equally operative in the human brain. Such constraints provide an evolutionarily plausible route to solving the "poverty of the stimulus problem" (aka Plato's problem: "how do we learn so much given so little input?") by pushing the answer much further back in evolutionary time than the recent split between humans and chimpanzees. This proposal has the very attractive empirical implication that most such constraints can be explored via experimental work with non-human animals, and their neural, developmental and ultimately genetic bases uncovered. This proposal does not exclude the possibility that humans have additional, species-unique capacities (e.g. in the social cognition domain: [91,217]), but it suggests that these are outweighed by more ancient and broadly shared capacities.

This proposal is directly analogous to the situation in morphology. The human propensity to develop hands with five fingers is "innate" (reliably developing and genetically grounded), but there is nothing species-specific about this: five-digit forelimbs are a general tetrapod trait going back to very early land vertebrates [37,203]. Like most aspects of human form, the biological roots of our hands run deep, and there is little evidence that our brains are much different. It is only the too-frequent conflation of "innate" with "species-specific" that obscures this likelihood, and makes my proposal about the shared nature of the biological constraints on language learning seem at all surprising.

Nonetheless, other vertebrates neither spontaneously develop language, nor acquire it from us with training. This suggests that, on top of a broad foundation of capacities shared with other species, humans have at least a few additional abilities (or propensities) that set us apart. Again, there is nothing unusual with observing that a species has unique traits: all species do. Nor does it seem controversial to suggest that humanity's most unusual traits are in the cognitive domain. Thus, although many of the biological pre-requisites of language acquisition are probably shared, computational capacities presumably remain that are uniquely well-developed in humans.

The "dendrophilia hypothesis" states that one of these core capacities is a novel (or greatly expanded) ability for tree-based computation. This tree-building capacity supplies the representational framework over which probabilities can be computed, using the more ancient Bayesian predictive framework that is widely shared. The hierarchicality of music suggests that the human propensity to process streams into trees is not unique to language; on the other hand little current evidence suggests that this propensity is shared with other primates (including chimpanzees). Thus, for example, the language-trained bonobo Kanzi uses and understands word order effectively (a feature visible at the surface level) but shows no evidence of deeper constraints involving phrase structure rules [191]. The productions of other language-trained chimpanzees, such as Nim, show little evidence for any rule-based structure at all [234]. Human children, by contrast, seem to start out with rather superficial representations, requiring no hidden abstract categories, but by the age of three show clear evidence of such abstractions as "noun phrase" and "verb" [10].

As noted in Section 2.3, contemporary computational solutions to language acquisition all involve computation over trees (typically generated using context-free grammars). From a Bayesian viewpoint, this entails a model which computes probability distributions over arbitrarily large collections of trees [109,163,211]; a system confined to computations over surface strings will not suffice. This is consonant with the founding insight of generative linguistics: that linguistic rules operate over hierarchical structures, and not over words *per se* [30]. Thus, the dendrophilic drive of humans to parse strings into trees, compute probability distributions over those trees, and then use these to creatively generate novel structures (and the attendant strings) seems to be a prerequisite for language acquisition, clearly present in our species and limited in others.

I am *not* claiming that trees are nowhere to be found in animal cognition. I strongly suspect that trees form an important *implicit* component of motor control, navigation and perhaps social cognition in other primates, and I continue to hope that tree structures may play some role in birdsong in at least some avian species. The difference, according to the dendrophilia hypothesis, is that we have neural circuitry that directly and flexibly encodes such tree structures,

in an abstract and sensory-domain-independent manner. Since most animals possess some ability to represent serial order over the leaves of a tree, our unusual capacity must be in encoding hierarchical level (branching depth). Although my colleagues and I have previously discussed the generation of new levels in terms of "recursion" [88], I now regret this, since the connotation of unbounded generation of arbitrary-depth trees is not a necessity for dendrophilia [56], but rather a mathematically concise way to get at what *is* a necessity, namely flexible tree-building allowing multiple nested hierarchical levels. This tree-building cognitive capacity may be less abstract and less powerful than full recursion, and potentially more empirically tractable. Note that neurally, it may be a relatively small evolutionary step to go from implicit trees, confined to particular cognitive and neural domains, to explicit, cross-domain trees (e.g. via the trans-cortical "scratchpad" model developed above in Section 6.1).

8. Overall conclusions

I conclude with a few comments about, and questions for, the future.

I have stressed the need for a computational approach to cognitive science that embraces a formally explicit and conceptually sophisticated theoretical framework. I observed that many long-running debates in cognitive science seem to be going nowhere because, lacking a clear and unambiguous theoretical framework capable of making testable quantitative predictions, they get mired in terminological morasses and in unhelpful dichotomies. Fortunately, there are abundant bodies of theory already available, in particular in single-neuron physiology, the theory of computation, and predictive systems theory, that fulfill many of the requirements for a rigorous, computationally- and biologically-grounded framework for cognitive science. As stressed initially, these islands of rigor discussed above by no means exhaust the store: game theory, network/graph theory, algorithmic complexity, and machine learning all provide important additional sources from outside "classical" cognitive science. Within cognitive science, music theory and linguistics both have pockets of rigorous theory ripe for connection to neural computation. Creating a framework incorporating bridging hypotheses to map between these various domains will allow cognitive biologists to generate and test explicit predictions about cross-species comparisons.

I have specifically highlighted the importance of tree-based computation in such a framework. At a high level this is because trees play a central and ineliminable role in human language and music, and thus provide one anchor at the cognitive end of the spectrum. At the other (microscopic) end, it is a simple fact that neurons have a tree-like form, and it has become clear in the last decades that this structure plays a central role in the cell-level computations they perform. Thus the form of individual neuronal computing elements, which must figure in any framework bridging the brain/mind gap, are also trees. It remains to be seen what, if any, link exists between the tree forms present at these cellular and cognitive poles, but I suspect that there are non-trivial connections between tree-based neural implementation and tree-oriented algorithms and computation. Finding out what they are will demand first acknowledging the facts of neuronal structure, and then exploring their potential implications for higher-level cognition at the algorithmic and computational levels. My personal hunch is that certain algorithms or representational formats will be more easily implemented in networks of tree-shaped computational elements than others, and therefore that at least some of the facts of cellular-level neuroscience will "percolate up" with implications for cognition and natural computation.

The goal of this overall endeavor is to someday create a field of computational cognitive biology able to answer questions like the following:

- Why are many birds able to achieve the same cognitive abilities as many mammals, but using much smaller brains?
- What specific neural changes enable some vertebrates, and not others, to be vocal learners?
- Why can only a small fraction of animal species find a rhythmic beat in music, and entrain their own movements to it?
- What specific neural differences between humans and chimpanzees enable one species and not the other to acquire complex linguistic syntax?

This is just a small selection of my personal favorite questions, ones that will probably keep me occupied for the rest of my scientific career. But many more questions in this vein are easily generated ("left as an exercise for the reader"). A mature field of cognitive science should be able to ask, and answer, this broad class of questions in explicit computational and neural terms.

Acknowledgements

I acknowledge the financial support of ERC Advanced Grant SOMACCA (Grant #230604), and thank Daniel Bowling and an anonymous reviewer for detailed comments on an earlier version of the manuscript.

References

[1] Abe K, Watanabe D. Songbirds possess the spontaneous ability to discriminate syntactic rules. Nat Neurosci 2011;14:1067–74.
[2] Aho A, Sethi R, Ullman JD. Compilers: principles, techniques and tools. Reading, Massachusetts: Addison-Wesley; 1986.
[3] Appel AW. Modern compiler implementation in C. Cambridge, UK: Cambridge University Press; 1998.
[4] Arbib MA, Caplan D. Neurolinguistics must be computational. Behav Brain Sci 1979;2:449–83.
[5] Ariew A. Innateness is canalization: in defense of a developmental account of innateness. In: Hardcastle VG, editor. Where biology meets psychology: philosophical essays. Cambridge, MA: MIT Press; 1999. p. 117–38.
[6] Atal BA, Hanauer SL. Speech analysis and synthesis by linear prediction of the speech wave. J Acoust Soc Am 1971;50:637–55.
[7] Bach E, Brown C, Marslen-Wilson W. Crossed and nested dependencies in German and Dutch: a psycholinguistic study. Lang Cogn Processes 1986;1:249–62.
[8] Bahlmann J, Schubotz RI, Friederici AD. Hierarchical artificial grammar processing engages Broca's area. Neuroimage 2008;42:525–34.
[9] Baldwin KB, Kutas M. An ERP analysis of implicit structured sequence learning. Psychophysiology 1997;34:74–86.
[10] Bannard C, Lieven E, Tomasello M. Modeling children's early grammatical knowledge. Proc Natl Acad Sci 2009;106:17284–9.
[11] Bates E, Elman J. Learning rediscovered. Science 1996;274:1849–50.
[12] Bateson P. The corpse of a wearisome debate. Science 2002;297:2212–3.
[13] Bateson P, Mameli M. The innate and the acquired: useful clusters or a residual distinction from folk biology? Dev Psychobiol 2007;49:818–31.
[14] Beckers GJL, Bolhuis JJ, Okanoya K, Berwick RC. Birdsong neurolinguistics: songbird context-free grammar claim is premature. NeuroReport 2012;23:139–45.
[15] Bell CC, Han V, Sawtell NB. Cerebellum-like structures and their implications for cerebellar function. Annu Rev Neurosci 2008;31:1–24.
[16] Berlinski D. The advent of the algorithm: the 300-year journey from an idea to the computer. San Diego: Harcourt; 2001.
[17] Berridge KC. The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl) 2007;191:391–431.
[18] Bharucha JJ. Music cognition and perceptual facilitation: a connectionist framework. Music Percept 1987;5:1–30.
[19] Blumstein SE. Phrenology, "boxology" and neurology. Behav Brain Sci 1979;2:460–1.
[20] Bookheimer S. Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. Annu Rev Neurosci 2002;25:151–88.
[21] Box GEP, Draper NR. Empirical model building and response surfaces. Oxford, UK: John Wiley & Sons; 1987.
[22] Breland K, Breland M. The misbehavior of organisms. Am Psychol 1961;16:681–4.
[23] Buckner C. Two approaches to the distinction between cognition and 'Mere Association'. Int J Comp Psychol 2011;24:314–48.
[24] Byrne RW, Bates LA. Why are animals cognitive. Curr Biol 2006;16:445–8.
[25] Caramazza A, Zurif EB. Dissociation of algorithmic and heuristic processes in language comprehension: evidence from aphasia. Brain Lang 1976;3:572–82.
[26] Cazé RD, Humphries M, Gutkin B. Passive dendrites enable single neurons to compute linearly non-separable functions. PLoS Comput Biol 2013;9:e1002867.
[27] Chater N, Manning CD. Probabilistic models of language processing and acquisition. Trends Cogn Sci 2006;10:335–44.
[28] Cheney DL, Seyfarth RM. Assessment of meaning and the detection of unreliable signals by vervet monkeys. Anim Behav 1988;36:477–86.
[29] Chomsky N. Three models for the description of language. IRE Trans Inf Theory 1956;IT-2:113–24.
[30] Chomsky N. Syntactic structures. The Hague: Mouton; 1957.
[31] Chomsky N. Aspects of the theory of syntax. Cambridge, Massachusetts: MIT Press; 1965.
[32] Chomsky N, Miller GA. Finite state languages. Inf Control 1958;1:91–112.
[33] Christiansen MH, Chater N. Toward a connectionist model of recursion in human linguistic performance. Cogn Sci 1999;23:157–205.
[34] Church A. A note on the Entscheidungsproblem. J Symb Log 1936;1:40–1.
[35] Churchland PS. The impact of neuroscience on philosophy. Neuron 2008;60:409–11.
[36] Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 2013;36:181–204.
[37] Daeschler EB, Shubin NH, Jenkins Jr. FA. A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. Nature 2006;440.
[38] Davis M. Computability and unsolvability. New York: McGraw-Hill; 1958.
[39] Davis M, editor. The undecidable. Basic papers on undecidable propositions, unsolvable problems and computable functions. Hewlett, NY: Raven Press; 1965.
[40] Daw ND, Frank MJ. Reinforcement learning and higher level cognition: introduction to the special issue. Cognition 2009;113:259–61.
[41] de Vries MH, Monaghan P, Knecht S, Zwitserlood P. Syntactic structure and artificial grammar learning: the learnability of embedded hierarchical structures. Cognition 2008;107:763–74.
[42] Dyson G. Turing's cathedral. New York: Pantheon; 2012.
[43] Eimas PD, Siqueland P, Jusczyk P, Vigorito J. Speech perception in infants. Science 1971;171:303–6.
[44] Elman JL, Bates E, Johnson MH, Karmiloff-Smith A, Parisi D, Plunkett K. Rethinking innateness: a connectionist perspective on development. Cambridge, MA: MIT Press; 1997.

[45] Emery NJ, Clayton NS. The mentality of crows: convergent evolution of intelligence in corvids and apes. Science 2004;306:1903–7.
[46] Enard W, et al. A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell 2009;137:961–71.
[47] Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, et al. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 2002;418:869–72.
[48] Evans N, Levinson SC. The myth of language universals: language diversity and its importance for cognitive science. Behav Brain Sci 2009;32:429–48.
[49] Everett DL. Cultural constraints on grammar and cognition in Pirahã. Curr Anthropol 2005;46:621–46.
[50] Feynman R. The character of physical law. Cambridge, Massachusetts: MIT Press; 1965.
[51] Fisher SE. Tangled webs: tracing the connections between genes and cognition. Cognition 2006;101:270–97.
[52] Fisher SE, Scharff C. FOXP2 as a molecular window into speech and language. Trends Genet 2009;25:166–77.
[53] Fitch WT. Nano-intentionality: a defense of intrinsic intentionality. Biol Philos 2008;23:157–77.
[54] Fitch WT. Prolegomena to a future science of biolinguistics. Biolinguistics 2009;3:283–320.
[55] Fitch WT. The evolution of language. Cambridge: Cambridge University Press; 2010.
[56] Fitch WT. Three meanings of "recursion": key distinctions for biolinguistics. In: Larson R, Déprez V, Yamakido H, editors. The evolution of human language: biolinguistic perspectives. Cambridge, UK: Cambridge University Press; 2010. p. 73–90.
[57] Fitch WT. The evolution of syntax: an exaptationist perspective. Front Evol Neurosci 2011;3:1–12.
[58] Fitch WT. Innateness and human language: a biological perspective. In: Tallerman M, Gibson KR, editors. The Oxford handbook of language evolution. Oxford: Oxford University Press; 2011. p. 143–56.
[59] Fitch WT. Unity and diversity in human language. Philos Trans R Soc Lond B, Biol Sci 2011;366:376–88.
[60] Fitch WT, Friederici AD. Artificial grammar learning meets formal language theory: an overview. Philos Trans R Soc Lond B, Biol Sci 2012;367:1933–55.
[61] Fitch WT, Friederici AD, Hagoort P. Pattern perception and computational complexity. Philos Trans R Soc Lond B, Biol Sci 2012;367:1925–32.
[62] Fitch WT, Hauser MD. Computational constraints on syntactic processing in a nonhuman primate. Science 2004;303:377–80.
[63] Fitch WT, Hauser MD, Chomsky N. The evolution of the language faculty: clarifications and implications. Cognition 2005;97:179–210.
[64] Fitch WT, Huber L, Bugnyar T. Social cognition and the evolution of language: constructing cognitive phylogenies. Neuron 2010;65:795–814.
[65] Fitch WT, Martins MD. Hierarchical processing in music, language and action: Lashley revisited. Ann NY Acad Sci 2014;1316:87–104.
[66] Fodor JA. Special sciences (or: the disunity of science as a working hypothesis). Synthese 1974;28:97–115.
[67] Fodor JA, Pylyshyn ZW. Connectionism and cognitive architecture: a critical analysis. Cognition 1988;28:3–71.
[68] Franklin DW, Wolpert DM. Computational mechanisms of sensorimotor control. Neuron 2011;72:425–42.
[69] Friederici AD. Pathways to language: fiber tracts in the human brain. Trends Cogn Sci 2009;13:175–81.
[70] Friederici AD. The brain basis of language processing: from structure to function. Physiol Rev 2011;91:1357–92.
[71] Friederici AD, Bahlmann J, Heim S, Schubotz RI, Anwander A. The brain differentiates human and non-human grammars: functional localization and structural connectivity. Proc Natl Acad Sci USA 2006;103:2458–63.
[72] Friederici AD, Pfeifer E, Hahne A. Event-related brain potentials during natural speech processing – effects of semantic, morphological and syntactic violations. Cogn Brain Res 1993;1:183–92.
[73] Friston KJ. Beyond phrenology: what can neuroimaging tell us about distributed circuitry? Annu Rev Neurosci 2002;25:221–50.
[74] Garcia J, Koelling RA. Relation of cue to consequences in avoidance learning. Psychon Sci 1966;4:123–4.
[75] Gardner H. The mind's new science: a history of the cognitive revolution. New York: Basic Books; 1985.
[76] Gazzaniga MS, Ivry RB, Mangun GR. Cognitive neuroscience: the biology of mind. New York: W. W. Norton; 1998.
[77] Gentner TQ, Fenn KM, Margoliash D, Nusbaum HC. Recursive syntactic pattern learning by songbirds. Nature 2006;440:1204–7.
[78] Gersting JL. Mathematical structures for computer science. New York: W H Freeman; 1999.
[79] Geschwind N. The organization of language and the brain. Science 1970;170:940–4.
[80] Gidon A, Segev I. Principles governing the operation of synaptic inhibition in dendrites. Neuron 2012;75:330–41.
[81] Gigerenzer G, Todd PM. Simple heuristics that make us smart. Oxford, UK: Oxford University Press; 1999.
[82] Giurfa M, Zhang S, Jenett A, Menzel R, Srinivasan MV. The concepts of 'sameness' and 'difference' in an insect. Nature 2001;410:930–3.
[83] Gribbin J. Science: a history. London: Penguin; 2002.
[84] Griebel U, Oller DK. Vocabulary learning in a Yorkshire terrier: slow mapping of spoken words. PLoS ONE 2012;7:e30182.
[85] Griffiths PE. What is innateness? Monist 2002;85:70–85.
[86] Hagoort P. Broca's complex as the unification space for language. In: Cutler A, editor. Twenty-first century psycholinguistics: four cornerstones. London: Lawrence Erlbaum; 2005. p. 157–72.
[87] Hagoort P. On Broca, brain, and binding: a new framework. Trends Cogn Sci 2005;9:416–23.
[88] Hauser M, Chomsky N, Fitch WT. The language faculty: what is it, who has it, and how did it evolve? Science 2002;298:1569–79.
[89] Heiligenberg W. Neural nets in electric fish. Cambridge, MA: MIT Press; 1991.
[90] von Helmholtz H. Handbuch der Physiologischen Optik. Hamburg: Voss; 1911.
[91] Herrmann E, Call J, Hernàndez-Lloreda MV, Hare B, Tomasello M. Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. Science 2007;317:1360–6.
[92] Herz A, Gollisch T, Machens CK, Jaeger D. Modeling single-neuron dynamics and computations: a balance of detail and abstraction. Science 2006;314:80–5.
[93] Heyes CM. Theory of mind in nonhuman primates. Behav Brain Sci 1998;21:101–34.
[94] Hickok G. The functional neuroanatomy of language. Phys Life Rev 2009;6:121–43.

[95] Hochmann J-R, Azadpour M, Mehler J. Do humans really learn A^nB^n artificial grammars from exemplars? Cogn Sci 2008;32:1021–36.
[96] Honing H. Without it no music: beat induction as a fundamental musical trait. Ann NY Acad Sci 2012;1252:85–91.
[97] Hopcroft JE, Motwani R, Ullman JD. Introduction to automata theory, languages and computation. Reading, Massachusetts: Addison-Wesley; 2000.
[98] Houghton G, Hartley T. Parallel models of serial behaviour: Lashley revisited. Psyche 1995;2.
[99] Huber DH, Wiesel TN. Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. J Neurophysiol 1965;28:229–89.
[100] Huber DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. J Physiol 1968;195:215–43.
[101] Insel TR, Shapiro LE. Oxytocin receptor distribution reflects social organization in monogamous and polygamous voles. Proc Natl Acad Sci USA 1992;89:5981–5.
[102] Insel TR, Wang ZX, Ferris CF. Patterns of brain vasopressin receptor distribution associated with social organization in microtine rodents. J Neurosci 1994;14:5381–92.
[103] Jackendoff R. Foundations of language. New York: Oxford University Press; 2002.
[104] Jackendoff R, Pinker S. The nature of the language faculty and its implications for evolution of language (reply to Fitch, Hauser, & Chomsky). Cognition 2005;97:211–25.
[105] Jäger G, Rogers J. Formal language theory: refining the Chomsky hierarchy. Philos Trans R Soc Lond B, Biol Sci 2012;267:1956–70.
[106] Jarvis ED. Brains and birdsong. In: Marler P, Slabbekoorn H, editors. Nature's music: the science of birdsong. 2004. p. 226–71.
[107] Jerison HJ. Evolution of the brain and intelligence. New York: Academic Press; 1973.
[108] Johnson M. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In: Proceedings of the 46th annual meeting of the association for computational linguistics: human language technologies. Columbus, OH: Association for Computational Linguistics; 2008. p. 398–406.
[109] Johnson M. Language acquisition as statistical inference. In: Anderson SR, Moeschler J, Reboul F, editors. The language–cognition interface. Geneva: Librarie Droz; 2013. p. 109–34.
[110] Johnson M, Riezler S. Statistical models of syntax learning and use. Cogn Sci 2002;26:239–53.
[111] Joshi AK, Vijay-Shanker K, Weir DJ. The convergence of mildly context-sensitive formalisms. In: Sells P, Shieber SM, Wasow T, editors. Processing of linguistic structure. Cambridge, MA: The MIT Press; 1991. p. 31–81.
[112] Kahneman D. Thinking fast and slow. New York: Farrar, Straus and Giroux; 2011.
[113] Kaminski J, Call J, Fischer J. Word learning in a domestic dog: evidence for 'fast mapping'. Science 2004;304:1682–3.
[114] Kleene SC. On notation for ordinal numbers. J Symb Log 1938;3:150–5.
[115] Kleene SC. Representation of events in nerve nets and finite automata. In: Shannon CE, McCarthy JJ, editors. Automata studies. Princeton: Princeton University Press; 1956. p. 3–40.
[116] Koch C. Computation and the single neuron. Nature 1997;385:207–10.
[117] Koch C. Biophysics of computation – information processing in single neurons. Oxford: Oxford University Press; 1998.
[118] Koch C, Segev I. The role of single neurons in information processing. Nat Neurosci 2000;3:1171–7.
[119] Koelsch S. Brain and music. London, UK: John Wiley & Sons; 2012.
[120] Koelsch S, Gunter TC, von Cramon DY, Zysset S, Lohmann G, Friederici AD. Bach speaks: a cortical "Language-Network" serves the processing of music. NeuroImage 2002;17:956–66.
[121] Koelsch S, Maess B, Friederici AD. Musical syntax is processed in the area of Broca: an MEG study. Neuroimage 2000;11:56.
[122] Koelsch S, Rohrmeier M, Torrecuso R, Jentschke S. Processing of hierarchical syntactic structure in music. Proc Natl Acad Sci 2013;110:15443–8.
[123] Krebs HA. The August Krogh principle: for many problems there is an animal on which it can be most conveniently studied. J Exp Zool 1975;194:221–6.
[124] Ladefoged P. Elements of acoustic phonetics. Chicago: University of Chicago Press; 1995.
[125] Larkum ME, Nevian T, Sandler M, Polsky A, Schiller J. Synaptic integration in tuft dendrites of layer 5 pyramidal neurons: a new unifying principle. Science 2009;325:756–60.
[126] Lashley K. The problem of serial order in behavior. In: Jeffress LA, editor. Cerebral mechanisms in behavior; the Hixon symposium. New York: Wiley; 1951. p. 112–46.
[127] Lerdahl F. Musical syntax and its relation to linguistic syntax. In: Arbib MA, editor. Language, music and the brain. Cambridge, Massachusetts: MIT Press; 2013. p. 257–72.
[128] Lerdahl F, Jackendoff R. A generative theory of tonal music. Cambridge, Massachusetts: MIT Press; 1983.
[129] Levelt WJM. Formal grammars in linguistics and psycholinguistics: volume 1: an introduction to the theory of formal languages and automata, volume 2: applications applications in linguistic theory, volume 3: psycholinguistic applications. The Hague: Mouton; 1974.
[130] Levelt WJM. Formal grammars in linguistics and psycholinguistics. Amsterdam: John Benjamins; 2008.
[131] Linz P. An introduction to formal languages and automata. Sudbury, Massachusetts: Jones & Bartlett; 2001.
[132] London M, Häusser M. Dendritic computation. Annu Rev Neurosci 2005;28:503–32.
[133] Longuet-Higgins HC. Artificial intelligence and musical cognition. Philos Trans R Soc Lond A 1994;349:103–13.
[134] Lorenz K. Evolution and modification of behavior. Chicago: University of Chicago Press; 1965.
[135] Maess B, Koelsch S, Gunter TC, Friederici AD. Musical syntax is processed in Broca's area: an MEG study. Nat Neurosci 2001;4:540–5.
[136] Marcus G. Startling starlings. Nature 2006;440:1204–7.
[137] Marcus GF, Vijayan S, Bandi Rao S, Vishton PM. Rule learning by seven-month-old infants. Science 1999;283:77–80.
[138] Markel JD, Gray AH. Linear prediction of speech. New York: Springer Verlag; 1976.
[139] Marler P. The instinct to learn. In: Carey S, Gelman R, editors. The epigenesis of mind: essays on biology and cognition. Hillsdale, NJ: Lawrence Erlbaum Associates; 1991. p. 37–66.

[140] Marr D. Vision: a computational investigation into the human representation and processing of visual information. San Francisco: W H Freeman & Co.; 1982.
[141] Martin JG. Rhythmic (hierarchical) versus serial structure in speech and other behavior. Pyschol Rev 1972;79:487–509.
[142] Matsumoto M, Takada M. Distinct representations of cognitive and motivational signals in midbrain dopamine neurons. Neuron 2013;79:1011–24.
[143] Matsunaga E, Okanoya K. Expression analysis of cadherins in the songbird brain: relationship to vocal system development. J Comp Neurol 2008;508:329–42.
[144] Matsunaga E, Okanoya K. Evolution and diversity in avian vocal system: an Evo-Devo model from the morphological and behavioral perspectives. Dev Growth Differ 2009;51:355–67.
[145] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33.
[146] McGraw LA, Young LJ. The prairie vole: an emerging model organism for understanding the social brain. Trends Neurosci 2009;33:103–9.
[147] Mermin ND. What's wrong with this pillow? Phys Today 1989;42:9–10.
[148] Meulders M. Helmholtz: from enlightenment to neuroscience. Cambridge, Massachusetts: MIT Press; 2010.
[149] Miller GA. Free recall of redundant strings of letters. J Exp Psychol 1958;56:485–91.
[150] Miller GA. Project grammarama. In: Miller GA, editor. Psychology of communication. New York: Basic Books; 1967.
[151] Miller GA. The cognitive revolution: a historical perspective. Trends Cogn Sci 2003;7:141–4.
[152] Miller GA, Chomsky N. Finitary models of language users. In: Luce RD, Bush RR, Galanter E, editors. Handbook of mathematical psychology. New York: John Wiley & Sons; 1963. p. 419–92.
[153] Minsky ML. Computation: finite and infinite machines. Englewood Cliffs, New Jersey: Prentice-Hall; 1967.
[154] Minsky ML, Papert SA. Perceptrons. Cambridge, Massachusetts: MIT Press; 1969.
[155] Mumford D. On the computational architecture of the neocortex: II the role of cortico-cortical loops. Biol Cybern 1992;66:241–51.
[156] Mumford D, Desolneux A. Pattern theory: the stochastic analysis of real-world signals. Natick, Massachusetts: A K Peters, Ltd.; 2010.
[157] Neville HJ, Nicol JL, Barss A, Forster KI, Garrett MF. Syntactically based sentence processing classes – evidence from event-related brain potentials. J Cogn Neurosci 1991;3:151–65.
[158] Nowak M, Komarova NL, Niyogi P. Evolution of universal grammar. Science 2001;291:114–8.
[159] Pallier C, Devauchelle A-D, Dehaene S. Cortical representation of the constituent structure of sentences. Proc Natl Acad Sci 2011;108:2522–7.
[160] Passingham RE. Broca's area and the origins of human vocal skill. Philos Trans R Soc Lond B, Biol Sci 1981;292:167–75.
[161] Patel AD. Language, music, syntax and the brain. Nat Neurosci 2003;6:674–81.
[162] Pemmaraju SV, Skiena SS. Computational discrete mathematics. New York: Cambridge University Press; 2003.
[163] Pereira F. Formal grammar and information theory: together again? Philos Trans R Soc Lond 2000;358:1239–53.
[164] Perfors A, Tenenbaum JB, Gibson E, Regier T. How recursive is language? A Bayesian exploration. In: van der Hulst H, editor. Recursion and human language. de Gruyter Mouton; 2010.
[165] Perfors A, Tenenbaum JB, Regier T. The learnability of abstract syntactic principles. Cognition 2011;118:306–38.
[166] Perruchet P, Rey A. Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? Psychon Bull Rev 2005;12:307–13.
[167] Pilley JW, Reid AK. Border collie comprehends object names as verbal referents. Behav Process 2011;86:184–95.
[168] Pinker S. Words and rules: the ingredients of language. Basic Books; 1999.
[169] Pinker S. The blank slate: the modern denial of human nature. 2002.
[170] Pinker S, Jackendoff R. The faculty of language: what's special about it? Cognition 2005;95:201–36.
[171] Poeppel D, Embick D. Defining the relation between linguistics and neuroscience. In: Cutler A, editor. Twenty-first century psycholinguistics: four cornerstones. London: Lawrence Erlbaum; 2005. p. 103–20.
[172] Poggio T. The Levels of Understanding framework, revised. Perception 2012;41:1017–23.
[173] Post EL. Recursively enumerable sets of positive integers and their decision problems. Bull Am Math Soc 1944;50:284–316.
[174] Pulvermüller F. Brain embodiment of syntax and grammar: discrete combinatorial mechanisms spelt out in neuronal circuits. Brain Lang 2010;112:167–79.
[175] Ramón y Cajal S. Texture of the nervous system of man and the vertebrates. Berlin: Springer; 1894/2004.
[176] Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 1999;2:79–87.
[177] Reber AS. Implicit learning of artificial grammars. J Verbal Learn Verbal Behav 1967;6:855–63.
[178] Reber AS. Implicit learning and tacit knowledge: an essay on the cognitive unconscious. Oxford, UK: Oxford University Press; 1996.
[179] Rensch B. Increase of learning capability with increase of brain-size. Am Nat 1956;15:81–95.
[180] Rescorla RA. Pavlovian conditioning: it's not what you think it is. Am Psychol 1988;43:151–60.
[181] Rey A, Perruchet P, Fagot J. Centre-embedded structures are a by-product of associative learning and working memory constraints: evidence from baboons (*Papio papio*). Cognition 2012;123:180–4.
[182] Rilling JK, Glasser MF, Preuss TM, Ma X, Zhao T, Hu X, et al. The evolution of the arcuate fasciculus revealed with comparative DTI. Nat Neurosci 2008;11(4):426–8.
[183] Rohrmeier M. Towards a generative syntax of tonal harmony. J Math Music 2011;5:35–53.
[184] Rolls ET, Deco G. Computational neuroscience of vision. Oxford: Oxford University Press; 2001.
[185] Rosenbaum DA, Cohen RG, Jax SA, Weiss DJ, van der Wel R. The problem of serial order in behavior: Lashley's legacy. Hum Mov Sci 2007;26:525–54.
[186] Rosenfeld R. Two decades of statistical language modeling: where do we go from here? Proc IEEE 2000;88:1270–8.

[187] Rummelhart DE, McClelland JL. Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations. Cambridge, Massachusetts: MIT Press; 1986.
[188] Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. Science 1996;274:1926–8.
[189] Samuels R. Innateness in cognitive science. Trends Cogn Sci 2004;8:136–41.
[190] Savage JE. Models of computation: exploring the power of computing. Reading, Massachusetts: Addison-Wesley; 1998.
[191] Savage-Rumbaugh ES, Murphy J, Sevcik RA, Brakke KE, Williams SL, Rumbaugh DM. Language comprehension in ape and child. Monogr Soc Res Child Dev 1993;58:1–221.
[192] Sawtell NB, Williams A, Bell CC. From sparks to spikes: information processing in the electrosensory systems of fish. Curr Opin Neurobiol 2005;15:437–43.
[193] Schenker NM, Hopkins WD, Spocter MA, Garrison AR, Stimpson CD, Erwin JM, et al. Broca's area homologue in chimpanzees (*Pan troglodytes*): probabilistic mapping, asymmetry and comparison to humans. Cereb Cortex 2010;20:730–42.
[194] Schultz W. Getting formal with dopamine and reward. Neuron 2002;36:241–63.
[195] Schultz W. Updating dopamine reward signals. Curr Opin Neurobiol 2013;23:229–38.
[196] Schultz W, Dickinson A. Neuronal coding of prediction errors. Annu Rev Neurosci 2000;23:473–500.
[197] Seidenberg MS, Elman JL. Do infants learn grammar with algebra or statistics? Science 1999;284:434–5.
[198] Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27(379):623–56.
[199] Shannon CE. Prediction and entropy of printed English. Bell Syst Tech J 1951;30:50–64.
[200] Shettleworth SJ. Stimulus relevance in the control of drinking and conditioned fear responses in domestic chicks (*Gallus gallus*). J Comp Physiol Psychol 1972;80:175–98.
[201] Shieber SM. Evidence against the context-freeness of natural language. Linguist Philos 1985;8:333–44.
[202] Shipp S. The importance of being agranular: a comparative account of visual and motor cortex. Philos Trans R Soc Lond 2005;360:797–814.
[203] Shubin N. Your inner fish: a journey into the 3.5 billion-year history of the human body. London: Penguin Books; 2008.
[204] Silver RA. Neuronal arithmetic. Nat Rev Neurosci 2010;11:474–89.
[205] Síma J, Orponen P. General-purpose computation with neural networks: a survey of complexity theoretic results. Neural Comput 2003;15:2727–78.
[206] Simon HA. The architecture of complexity. Proc Am Philos Soc 1962;106:467–82.
[207] Skiena SS. The algorithm design manual. New York: Springer Verlag; 1998.
[208] Skinner BF. Verbal behavior. New York, NY: Appleton-Century-Crofts; 1957.
[209] Srinivasan MV, Laughlin SB, Dubs A. Predictive coding: a fresh view of inhibition in the retina. Proc R Soc Lond B, Biol Sci 1982;216:427–59.
[210] Stabler EP. Varieties of crossing dependencies: structure dependence and mild context sensitivity. Cogn Sci 2004;28:699–720.
[211] Steedman M. Romantics and revolutionaries. Linguist Issues Lang Technol 2011;6:1–20.
[212] Steedman MJ. A generative grammar for jazz chord sequences. Music Percept 1984;2:52–77.
[213] Stobbe N, Westphal-Fitch G, Aust U, Fitch WT. Visual artificial grammar learning: comparative research on humans, kea (*Nestor notabilis*) and pigeons (*Columba livia*). Philos Trans R Soc Lond B, Biol Sci 2012;367:1995–2006.
[214] ten Cate C, Okanoya K. Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. Philos Trans R Soc Lond B, Biol Sci 2012;367:1984–94.
[215] Tervaniemi M. Musical sound processing: EEG and MEG evidence. In: Peretz I, Zatorre RJ, editors. The cognitive neuroscience of music. Oxford: Oxford U. Press; 2003. p. 294–309.
[216] Tomalin M. The formal origins of syntactic theory. Lingua 2002;112:827–48.
[217] Tomasello M. The cultural origins of human cognition. Cambridge, Massachusetts: Harvard University Press; 1999.
[218] Trappenberg TP. Fundamentals of computational neuroscience. Oxford: Oxford University Press; 2002.
[219] Turing AM. On computable numbers, with an application to the Entscheidungsproblem. Proc Lond Math Soc 1937;42:230–65.
[220] Uddén J, Ingvar M, Hagoort P, Petersson KM. Implicit acquisition of grammars with crossed and nested non-adjacent dependencies: investigating the push-down stack model. Cogn Sci 2012;2012:1–24.
[221] Ujfalussy B, Kiss T, Érdi P. Parallel computational subunits in dentate granule cells generate multiple place fields. PLoS Comput Biol 2009;5:e1000500.
[222] Valiant L. Probably approximately correct: nature's algorithms for learning and prospering in a complex world. New York: Basic Books; 2013.
[223] van Heijningen CAA, de Vissera J, Zuidema W, ten Cate C. Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. Proc Natl Acad Sci 2009;106:20538–43.
[224] Vargha-Khadem F, Gadian DG, Copp A, Mishkin M. FOXP2 and the neuroanatomy of speech and language. Nat Rev Neurosci 2005;6:131–8.
[225] Vargha-Khadem F, Watkins K, Price CJ, Ashburner J, Alcock K, Connelly A, et al. Neural basis of an inherited speech and language disorder. Proc Natl Acad Sci USA 1998;95:12695–700.
[226] Vernes SC, Oliver PL, Spiteri E, Lockstone HE, Puliyadi R, Taylor JM, et al. Foxp2 regulates gene networks implicated in neurite outgrowth in the developing brain. PLoS Genet 2011;7:e1002145.
[227] Vijay-Shanker K, Weir DJ. The equivalence of four extensions of context-free grammars. Math Syst Theory 1994;27:511–46.
[228] Vuust P, Ostergaard L, Pallesen KJ, Bailey C, Roepstorff A. Predictive coding of music – brain responses to rhythmic incongruity. Cortex 2009;45:80–92.
[229] Wacongne C, Changeux J-P, Dehaene S. A neuronal model of predictive coding accounting for the mismatch negativity. J Neurosci 2012;32:3665–78.
[230] Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S. Evidence for a hierarchy of predictions and prediction errors in human cortex. Proc Natl Acad Sci 2011;108:20754–9.

[231] Westphal-Fitch G, Huber L, Gómez JC, Fitch WT. Production and perception rules underlying visual patterns: effects of symmetry and hierarchy. Philos Trans R Soc Lond B, Biol Sci 2012;367:2007–22.
[232] Wild JM. The avian nucleus retroambigualis: a nucleus for breathing, singing and calling. Brain Res 1993;606:119–24.
[233] Wolfe JM. What can 1 million trials tell us about visual search? Psychol Sci 1998;9:33–9.
[234] Yang C. Ontogeny and phylogeny of language. Proc Natl Acad Sci USA 2013;110:6323–7.