From “The Neocortex,” edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

17

Computation and Its Neural Implementation in Human Cognition

Lucia Melloni, Elizabeth A. Buffalo,
Stanislas Dehaene, Karl J. Friston, Asif A. Ghazanfar,
Anne-Lise Giraud, Scott T. Grafton, Saskia Haegens,
Bijan Pesaran, Christopher I. Petkov,
and David Poeppel

Abstract

How do the computations of the cerebral cortex and subcortical structures account for human perception, cognition, and affect? Answering this question requires understanding how the neurobiological and functional properties of the human brain give rise to the repertoire of human faculties and behavior, and hence, an understanding of the *neural mechanisms* that implement these functions. While research over the past decades has made substantial progress toward this end, significant challenges still lie ahead, and new opportunities open up daily as neuroscience and related fields develop and implement new theories and technologies. To (begin to) address these challenges, this chapter explores conceptual and methodological aspects inherent to the study of the neurobiology of the human mind that are at the core of the current "central paradigm" (Kuhn 1962) in neuroscience, but are often taken for granted and undergo little scrutiny. In particular, it discusses what defines or constitutes "uniquely human" mental capacities, the promises and pitfalls of using animal models to understand the human brain, whether neural solutions and computations are shared across species or repurposed for potentially uniquely human capacities, and what inspiration and information can be drawn from recent developments in artificial intelligence. Attention is given to laying out desiderata for future investigations into the human mind.

Group photos (top left to bottom right) Lucia Melloni, David Poeppel, Bijan Pesaran, Karl Friston, Elizabeth Buffalo, Scott Grafton, Asif Ghazanfar, Stan Dehaene, Anne-Lise Giraud, Lucia Melloni, Christopher Petkov, Saskia Haegens, Bijan Pesaran, Scott Grafton, Elizabeth Buffalo, Christopher Petkov, Stan Dehaene, Anne-Lise Giraud, Asif Ghazanfar, David Poeppel, Karl Friston

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

Singular or Not? The Human Animal

To understand the structure and function of *human* brains, one must address the question of potential uniqueness or singularity, ever cognizant that the word "uniqueness" raises its own set of provocative questions. The human cortex has structural and physiological properties that underwrite neuronal activity which in turn underpins the implementation of computations that may or may not be specific to the function of the human brain. Further definition, however, is required when we ask whether the potential distinctions between human brain, computational inventory, or behavioral repertoire, compared to other species, are a matter of degree, as argued for by Darwin (1888), or systemic discontinuities (Fitch 2012; Parravicini and Pievani 2018; Ghazanfar, this volume). Discontinuity of evolution is clearly not an idea that is widely endorsed. Rather than rehearse potential uniqueness features, we might instead pursue the argument of commonalities. We contend, however, that this is just as difficult as identifying properties that are apparently found only in human cortex.

For the sake of argument, we take the position that some distinctions between humans and other species can be readily identified, and we take it as our task to understand how to account for such species-typical features. One example that points to human-specific organization concerns the suite of operations that comprise *combinatorics*. These come to light in language, mathematics, music, theory of mind, and potentially in other domains not yet understood as well in terms of formalization.

In the language domain, it has long been argued that only humans have the capacity to produce the kinds of representations characteristic of syntax (e.g., Merge). To date, there is no clear case of a nonhuman primate that has learned to combine words systematically according to a complex grammar (Terrace et al. 1979; Yang 2013). In the few cases in which nonhuman primates have been able to produce sequences adhering to a supra-regular grammar, this was only accomplished after extensive training (over 10,000 trials), whereas preschool children master this behavior in less than five trials (Wang et al. 2018). In numerical cognition (i.e., the sense of number and capacity for mathematical thinking), it is well established that monkeys and humans start life with a similar approximate number system (Dehaene 2011). The acquisition of verbal counting and a system of Arabic numerals allows human children to move from an approximate, compressive representation of numerical quantities to an exact, linear system of number (Siegler and Opfer 2003; Dehaene 2011). In the absence of formal education (e.g., in indigenous Amazon populations), the approximate system remains largely unchanged in human adults (Pica et al. 2004). Monkeys can be taught some number symbols, and this leads them to become somewhat more precise in a number comparison task (Livingstone et al. 2010). They may even begin to understand the rudiments of addition and subtraction (Livingstone et al. 2014). Yet, they continually make errors, even under highly motivating reinforcement schedules, and never perform at

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

the level of precision and exactness attained even by young human children. Arguably, sharp distinctions between precise consecutive numbers, as those between truth and falsehood, may be unique to humans: the formation of a complex combinatorial system of arithmetic certainly is.

Accepting that there are domains of behavior that may be singular to humans leads us to ask in which ways those operations are different: Are they rooted in simpler forms of behavior that can be useful to study when trying to understand how uniquely human behaviors emerged? In which way are the structure of the cerebral cortex and its neural codes different, relative to the brains and codes that implement these more basic behaviors? Have neural codes been exapted or created *de novo* for new functions?

Addressing these questions is inextricably linked with exploring neural solutions in other species to establish convergence and divergence. The usefulness of a comparative approach to understand the human brain and its dysfunctions is clear, yet there are a number of outstanding challenges that complicate such an enterprise. These challenges must be factored into any discussion if progress is to be made in understanding which neural solutions and computations might be shared across species or repurposed for potentially uniquely human capacities.

The Challenges of Understanding the Neurobiology of Human Cognition through Animal Models

The most straightforward approach to understanding the complexities of the human brain is to study the human brain itself, and with the emergence and refinement of a range of neuroimaging technologies, progress has been achieved over the last decades. For ethical and technological reasons, however, *direct* access to neural activity on a spatial scale, deemed necessary to unravel the neural computations that give rise to cognition and perception (i.e., for populations of individually resolvable neurons), is extremely limited in humans. Such recordings are currently only possible in patients who undergo brain surgery for tumor resection (e.g., Desmurget et al. 2009), implantation of deep-brain stimulation electrodes (e.g., Wahl et al. 2008; Cavanagh et al. 2011), or invasive epilepsy monitoring (e.g., Ding et al. 2016; Schwiedrzik et al. 2018); that is, only in brains that are affected by disease. These recordings are serendipitous in nature because they rely on recording sites that are selected for monitoring based solely on clinical considerations. Therefore, we must rely on animal models for the most part to leverage neuroscientific toolsets available for the study of the brain, including system perturbation and circuit manipulations. This necessitates making assumptions and compromises about the aspects of human cognition that can be realistically modeled by the species serving as a model.

One common implicit and often untested assumption in the field pertains to *homology*. Specifically, it is commonly assumed that differences among

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

species are a matter of degree, such that mechanisms are conserved across phylogeny (perhaps with a scaling function). Thus, studies in other animals are thought to advance our understanding of human cognition and the human brain through a relatively straightforward translation of findings from one brain to another. Taking nonhuman animals as a sufficiently faithful model of human brain function or dysfunction (e.g., for depression or schizophrenia) is clearly useful and has provided important insight into the neurobiology of cognition. For example, although their evolutionary lineages split some 25 million years ago (Kumar and Hedges 1998), Old World nonhuman primates and humans both have a system to represent numerical quantities with striking similarities (Nieder and Dehaene 2009). At the same time, since we do not fully understand mammalian and cross-species homologies, we are often surprised at how challenging it is to translate pathophysiological mechanisms from murine animal models to primate models to humans for clinical purposes (Sena et al. 2010; van der Worp et al. 2010). This leads to questions on how readily insights in basic neuroscience obtained in another species are translatable to humans without a more complete understanding of homologies and specializations across the relevant species. Furthermore, as our understanding of human and nonhuman brains advances, more differences become apparent: the organizational principles of inter-areal connections, for example, seem to differ fundamentally between rodents and primates (Horvat et al. 2016; Gamanut et al. 2018), and there are multiple, potentially nontrivial differences in the structure of the visual system between macaque monkeys and humans (Preuss 2004). Still, commonalities also become more evident: even parts of prefrontal cortex thought to have specialized in humans, such as Broca's area, show remarkably conserved cyto- and receptor-architectonic patterns between monkeys, apes, and humans (Zilles and Amunts 2018).

The mere issue of establishing homology is complicated in and of itself (Rendall and Di Fiore 2007; Hall 2013). In the past, homology has predominantly been addressed on the level of morphological features (structural homology). Nowadays, the concept of homology has been expanded to other aspects, including behavior (phenotypical homology). However, as yet there is no consensus as to the level (neural, computational, algorithmic) needed for homologies to be useful or the criteria (e.g., genetic, developmental) required to constitute evidence for homology to distinguish it, for example, from analogy. In addition, phenotypical features can be superficially similar but of separate evolutionary origin because they have both experienced similar selective pressure. Thus, to test rigorously for homologies or identify the form of specialization, it seems pertinent to gather evidence from behavior, genetics, development, *and* neurobiological mechanisms. In addition, neurobiological evidence should be gathered across several levels (e.g., architectonic, morphological, neurophysiological).

As pointed out by Ghanzanfar (this volume), differences and commonalities between species should not only consider the brain but also the body, *and*

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

sensorium. For example, it was widely accepted that nonhuman primates do not speak due to vocal limitations in the anatomy and the configuration of their vocal tract. Detailed X-ray studies of the vocal tract in nonhuman primates, however, led to a rejection of this hypothesis: the vocal production apparatus in primates is capable of producing five clearly distinguishable vowels (Fitch et al. 2016; Boë et al. 2017). Differences must therefore lie elsewhere and remain to be identified. This example clearly demonstrates the need to consider differences at several levels to draw firm conclusions about whether a set of behaviors is similar or different across species.

We note that similar to the problems inherent to a consideration, for example, of just the brain, there are limitations in relying on observed behavior alone to infer species uniqueness or, more importantly, non-uniqueness. One problem concerns multiple realizability. As mentioned above, superficial similarity does not guarantee shared evolutionary origins: the same behavior in two species can be due to profoundly different, underlying cognitive operations and neural mechanisms. Even if we focus solely on human behavior, there are many classic cases of multiple realizations in sequence processing (Grafton et al. 1995; Schendan et al. 2003), visual category, and procedural learning (Clower and Boussaoud 2000). For example, in visual category learning, a subject learns through feedback whether stimuli are members of one category or another. Critically, a subject can draw on at least two learning mechanisms to develop the skill (Ashby and O'Brien 2005). Depending on the literature, one mechanism is referred to as reinforcement learning, procedural learning, implicit learning, model-free learning, or information integration. The other mechanism is referred to as rule-based, explicit, or model-based learning. Both mechanisms draw on different neural circuits (roughly, dopamine/striatal mediated and cortical) during training, and final performance is dependent on different neural systems. Whether or not a given species will draw on each of these learning mechanisms is highly dependent on brain design and task complexity (Smith et al. 2012a). These distinctions cannot be formed by observing behavioral performance in isolation. The ambiguity of multiple realizations necessitates additional evidence via task decomposition, ontological approaches, or neurophysiological methods as well as approaches from artificial intelligence (AI). This leads to a reframing of the question to one that examines species-specific *means* for accomplishing a given behavior rather than one of uniqueness in any given species (Smith et al. 2012b).

Another challenge to consider is whether our experimental assays preclude us from seeing similarities between species (i.e., a Type 2 problem). Failure to detect relevant similarities may result from the fact that we are forcing experimental animals to execute tasks that are not part of their natural repertoire; if so, we would expect behavior to be optimized for their own species-typical learning apparatus (Krakauer et al. 2017). An alternative approach would be to use evolutionarily more remote species for specific traits that they may or may not share with us, rather than using monkeys (the closest available model

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

system to the human brain) as a proxy for human behavior and cognitive functioning. In the language domain, we have tried to address certain aspects of language in apes and monkeys with limited success, specifically when it comes to higher-order combinatorics. For instance, chimpanzees using signs could not combine more than two symbols for communication (Terrace et al. 1979; Yang 2013). Importantly, there has been no systematic coevolution between monkeys and humans; monkeys have developed their own communication system, which does not possess key features of human language. There are species, however, that have coevolved with humans, that are under heavy pressure to understand human language, and which have developed speech perception skills (Andics et al. 2016). Dogs, in particular, are exposed to human language from birth and yet never acquire the ability to produce speech. Dogs do understand human orders made by specific word sequences (Bloom 2004; Kaminski et al. 2004; Pilley and Reid 2011). These types of animal models can serve to address questions about language processing in the human brain; for example, how much the speech production system contributes to speech perception and whether combinatorial properties are specifically human. Other remote species which have not coevolved with humans, yet show similar levels of encephalization as apes and humans, and have had specific pressures (unlike apes) to communicate by the auditory modality (e.g., cetaceans), may also be useful to study. Despite the absence of coevolution with humans, dolphins are able to understand word sequences (Herman and Morrel-Samuels 1995), which might mean that they also use temporally structured sequences of abstract symbols in their own cognitive functioning. These highly adapted mammals rely entirely on oral communication to maintain contact with their offspring and hence represent yet another alternative model of complex oral communication.

Another alternative is to explore repertoires of behaviors that animals exhibit in the wild as these may offer structural similarities to the computation under scrutiny. For instance, to understand whether recursivity is a feature exhibited in other animals, a potentially fruitful approximation of how animals establish hierarchies, even an atypical one (e.g., center-embedded dependencies), would be worth exploring. Work by Cheney and Seyfarth's group (e.g., Bergman et al. 2003) shows that baboons use their knowledge of social dominance hierarchies to evaluate vocal exchanges between animals with different social rank. Other work in the visual domain suggests that human infants evaluate object shape and color hierarchically (Werchan et al. 2015). Paradigms such as these provide a glimpse into combinatorial operations that respect certain hierarchical dependencies. In nonhuman animals, hierarchical dependencies may not capture the full complexity of the problem (e.g., manipulations of word classes in relation to syntactic knowledge) but they might permit us to get at the core neurobiological processes that support various aspects of these operations. The main advantage of this approach is that it builds on sets of behaviors and operations for which animals have evolved, that they naturally exhibit, thus

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

taking us away from artificial paradigms which may constrain, and possibly misguide, the results that we get. One might argue that the approach somehow preempts the answer, as the implicit assumption is that the natural behavior is already a good approximation to the computation that we are trying to test in humans. This is not necessarily the case if one couples this approach with further tests that constrain the problem. For instance, one could test whether the neural instantiation of the specific operation under scrutiny is the same between humans and animals through neuroimaging and/or electrophysiology (Tsao et al. 2008; Yovel and Freiwald 2013; Schwiedrzik et al. 2015; Wilson et al. 2015; Kikuchi et al. 2017; Sliwa and Freiwald 2017).

Taken together, it seems worthwhile to reconsider how "uniqueness" is defined and how to determine whether discontinuities exist in evolution. Potential pitfalls in the current research program include negligence of non-brain aspects in the assessment of similarities and differences between animal models and humans as well as an overreliance on behavior alone. Finally, a potentially fruitful avenue is to expand the range of available model systems, specifically targeting animals that have evolved circumscribed capabilities that may help us understand aspects of functions, such as language, that we consider uniquely human, as well as tapping onto natural behaviors that animals exhibit in the wild as opposed to employing artificial tasks as is currently done.

Notwithstanding the challenges, countless examples have already demonstrated the usefulness of animal models in illuminating the human brain and its dysfunction: the discovery of spatial codes in rodent medial temporal structures by John O'Keefe, May-Britt Moser, and Edvard I. Moser has direct relevance on our understanding of human cognition (O'Keefe 1976; Fyhn et al. 2004); studies by Benabid and DeLong in monkeys paved the pathway for deep brain stimulation treatment in Parkinson patients (Benabid et al. 1991; Bergman et al. 1994); and the interdisciplinary work by Peter Dayan, Ray Dolan, and Wolfram Schultz in human and nonhuman primates identified the neural computations for reward-related learning with implications for addiction, gambling, and clinically impaired decision making (e.g., Schultz 2015). Below, we explore how insights gathered from animal models *can* help us understand the human brain, and its potential unique set of cognitive operations.

Repurposing the Old: From Sequences to Combinatorics

The issue of uniqueness also arises at the level of basic neural computations. Is there a set of common neural computations across species? If so, can this set explain aspects of human cognition that are putatively unique? Or are there neural codes that are themselves uniquely human?

To begin to address these questions, we take the case of language, as we think it offers fruitful starting points for discussion in relation to computations

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

that might be shared across species versus computations which might be an attribute of human cortex. Some of the best-supported evidence for neural codes comes from sensory domains. However, insight derived from sensory modalities (while relevant) does not currently offer an explanation (or even a satisfying clarification) of the problem of linguistic representation and computation. Specific neural computation "for language" must be examined at a level of abstraction that goes beyond sensory and motor coding, because linguistic representation and computation can stem from auditory (speech), visual (sign language, text), and somatosensory information (Braille). In all these cases, the sensory modalities provide interface information to linguistic computations, which have specific, abstract properties. Figure 17.1 schematizes the nature of the problem: sequential information is processed by the sensorimotor interfaces (the input and output strings), but the system must be able to traffic in structured representations that permit computations over representations that go well beyond linear strings.

To be sure, there are other aspects of perception and cognition that may capitalize on some of the operations on hierarchies that we discuss here, notably action planning and movement, spatial navigation, and visual scene analysis. However, we focus on language because cross-species work is particularly complicated. Aspects of language that merit explanation include the property of discrete infinity, (nested) hierarchy, structure dependence, constituency, and the organization of the mental lexicon. To operationalize these key concepts

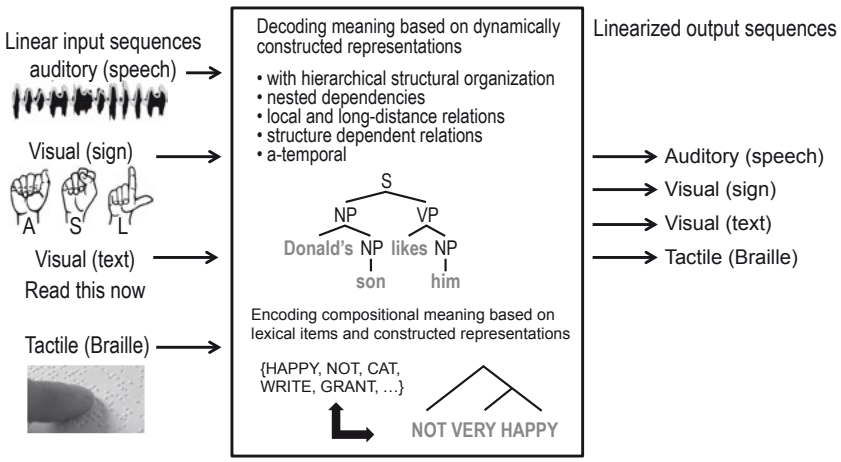

Figure 17.1 Neural computation "for language" requires abstract coding schemes. Sequential and linear information processed by the sensorimotor interfaces, the input and output strings, must be transformed into structured, hierarchical representations, allowing for computations of representations that go well beyond linear strings (e.g., recursivity).

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

and provide examples to work through the challenges, we characterize the issues as follows:

1. There are terminal (basic) elements, roughly words (e.g., the word "lock"). Even within words, there is compositional structure: strings (morphemes) can be concatenated to create different words, which, in turn, depending on the specific form of concatenation, result in different meanings due to structural ambiguity (e.g., "un-lockable" versus "unlock-able").
2. Words can be combined (e.g., "red boat" or "bad example") such that the resultant item inherits the properties of just one of the elements (a subroutine sometimes called labeling). That is, a "bad example" is a constituent, and this constituent bears the label "type of example" but not "type of bad."
3. The concatenation of words is based on structural (syntactic) and meaning-based (semantic) constraints: "new plans give hope" is parsed into the constituents [new plans] [give hope].
4. Recursivity: In the phrase "fast red boat," "red" modifies "boat" and "fast" modifies "red boat." This is an example of the recursive application of a rule, in which first the modifier A ("red") is applied to object B ("boat") to yield a new object, B′: [B′ /\ A B].

We advance the hypothesis that understanding such a generative system requires breaking the problem into formal operations (computations) that comprise the system. Those formal operations might map onto specific neural responses that may be amenable to neural coding research. As a starting point, we take the taxonomy of sequential operations illustrated by Stanislas Dehaene (see Figure 15.1, this volume): transitions and timing knowledge, chunking, ordinal knowledge, algebraic patterns, and nested tree structures.

Evidence shows that the first four levels represent sequence construction operations and sequence representations that are shared with other animals; in contrast, current evidence points to the conjecture that combinatorics which yield hierarchical nested tree structures might be a human singularity (for further discussion, see Dehaene, this volume). According to this hypothesis, what makes human thought complex (and, perhaps, of a certain kind) is that symbols (in language, mathematics, and perhaps music, action planning/motor control, and visual scene understanding) are not just strung together into a sequence (string-of-beads hypothesis); they are mentally represented as hierarchically structured trees (Calder-mobile hypothesis), thus offering combinatorial and interpretive diversity. This raises the question whether existing neural codes[1] might be used to represent such sequences and, if so, whether they are

[1] We explicitly do not concentrate on the question of what the neural code might mean: whether neurons really represent an external property or not, and what the neural code could be (e.g., rate code, state-space trajectories).

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

implemented similarly across different species. If the processing of tree structures that impose structure *dependence* is *not* shared across animals, it is still pertinent to determine whether neural codes have been exapted in evolution or created *de novo* for such purpose.

To get traction on the problem, we suggest reducing it to the establishment and representation of *relations*. The argumentation strategy we pursue here is to turn to the implementational level of description, in the sense of Marr (1982), looking to neurobiological properties that may motivate research on relations as they might be described at the algorithmic and computational levels. The desideratum for the language case is that the formal relation can express hierarchical nesting. The need to represent relations is also present for many other domains, including visual scene analysis, action planning, and motor control.

In vision, in particular for scene perception (and, more challengingly, scene understanding), relations need to be established: from exploring with our eyes to forming a scene representation to using it, say, to grasp an object. This includes the representation of spatial and topological relations: a pair of glasses *on* top of the table to the *left* of the cup. In even the simplest kinds of motor action, such as picking up the glasses on the table, a relation is formed through the intimate timing between the velocity of the arm as it reaches toward the object and the opening of the fingers which achieve a maximum aperture at a highly reproducible moment before enclosing the glasses (Paulignan et al. 1990). At first glance, this hierarchical relation is dominated by the kind of grasp the object requires, which regulates the arm speed. If, however, the cup is in the way of the glasses, then the hierarchical relation flips, with the limb trajectory dominating the timing of subsequent events. As simple and intuitive as these examples might seem, we do not fully understand how this is accomplished.

To further illustrate how considering the representation of relations in other domains can inform research on language, we explore here the prima facie similarity with spatial navigation in more detail. In spatial navigation, relations need to be established from moving around to forming a map to using it to navigate. Navigation requires chaining operations, or a series of sequences to find a path from A to B. Considering that superficial similarity, we turn to neural codes observed in the hippocampus. As rats run through a maze or forage in an open field, place cells in the hippocampus create a representation of the animal's environment (O'Keefe 1976), and ensembles of place cells fire in ordinal sequences that reflect the rat's ongoing experience (Dragoi and Buzsáki 2006). The hippocampal local field potential exhibits a prominent theta band (6–10 Hz) oscillation as the rat explores and actively processes incoming information. Importantly, within a theta cycle, the temporal offset between sequentially firing neurons is tightly correlated with the distance between each neuron's place field (Geisler et al. 2007), and these "theta sequences" incrementally advance across progressive theta cycles (Dragoi and Buzsáki 2006). These features of sequences within theta cycles allow the population of hippocampal neurons to

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

link temporally disparate events with a sequentially active ensemble. That is, within each theta cycle, the sequential firing of place cells provides a representation of the rat's previous, current, and future location, thus providing a way to tile the gaps between experienced events (Buzsáki and Llinas 2017).

It is conceivable that these aspects of hippocampal activity map on to at least the first three stages of the sequences described by Dehaene et al. (2015), with the sequential firing of place cells reflecting the transition and timing of sensory experiences as the rat runs through a maze (Stage 1), and the repetition of sequences within a theta cycle reflecting both chunking (Stage 2) and ordinal knowledge (Stage 3). Interestingly, hippocampal sequences can be replayed forward as well as backward, and there is some evidence that the forward sweeps may reflect prospective coding whereas backward sweeps reflect retrospective coding (Diba and Buzsáki 2007), both of which are thought to support mechanisms of episodic memory. Theta sequences reflect ordinal knowledge (in conjunction with specific item features) because the timescale of the sequential replay of activity within a theta cycle is independent of the timescale of experienced events; the sequence, for instance, maintains only the relative temporal order of experienced places.

The codes described in the hippocampus could serve as pointers for further research in cortex. We note that the role of the hippocampus in language, music, or mathematics is poorly understood. Recent studies, however, have suggested that the hippocampus might play a role in language (Piai et al. 2016) as well as in statistical learning (Schapiro et al. 2014, 2017). The latter is thought to be a mechanism guiding the discoveries of words in continuous speech (Saffran and Kirkham 2018). Here, the question is how babies discover, parse, segment, and string units for further processing in the continuous acoustic stream. Neural responses observed in the hippocampus (e.g., theta phase precession, replay, pattern completion, and/or pattern separation) may provide starting points to understand how primitive sequential operations are implemented in the human brain. Still, although these neural processes have advanced our understanding of sequence representation, they only represent sequences as temporal successions of events (i.e., the string-of-beads hypothesis described above) and not as fully abstract, a-temporal, hierarchical representations, such as those observed in language and mathematics (i.e., the Calder-mobile hypothesis). Hence, further refinement of our understanding of those codes is needed to explain hierarchical relations.

Other potential mechanisms for the implementation of sequence processing can be considered. Below, we briefly outline a number of possible candidates for the implementation of different sequential operations.

Chunking Operations

Here, two mechanisms are germane. The first is implemented through anatomical convergence. Feature-sensitive nodes A and B converge on

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

conjunction-specific node C, which after appropriate adjustment of synaptic gain and thresholds will respond selectively to conjunctions of feature A and B. This strategy is commonly used in hierarchically structured feedforward networks, including deep convolutional networks. The wiring can either be genetically determined (e.g., motion detectors) or specified by experience, using an associative Hebbian mechanism of synaptic plasticity. This results in the implementation of a conjunction-specific node that reflects frequently occurring statistical contingencies of features. While this mechanism is extremely robust to establish learned and stable patterns of relations, an open question is whether and how such a mechanism could be extended to *online* sequence construction; that is, constructing a chunk or "type" based on sparse data. Another question is whether a convergence site "C" is even needed. It could be that the convergence sites merely hold a combinatorial code linking information available in representations A and B, wherever they may be held (Damasio 1989).

The second mechanism is the formation of Hebbian assemblies, via recurrent activity, consisting of reciprocally coupled nodes that respond to different features. Again, through Hebbian learning, coupling connections among the nodes of the future assembly will be strengthened such that those nodes will be coupled preferentially to represent frequently co-occurring sets of features. As a result, if the corresponding set of features is present, the assembly representing the conjunction will be ignited. This strategy requires recurrence, a typical property for cortex, and because recurrency generates additional dynamics it can also be used to associate (chunk) more complex features, such as particular sequences. Thereby, hippocampal recurrent activity could help to bind items A and B and make their sensory cortical neural representations sparser and more similar (Messinger et al. 2001).

Establishment of Sequence Order

Recurrent networks reverberate and are self-active as well as generative. They have fading memory (stimuli leave long-lasting traces in reverberating activity) and can therefore integrate (chunk) responses evoked by sequentially presented stimuli. If a node in such a network is activated, it produces "songs," (i.e., sequences of successively activated nodes), whereby the sequence depends on the functional architecture of the reciprocal coupling connections. As their weight distribution reflects statistical contingencies of previous input (experience), the "songs" correspond to the encoding of learned sequences. These can then be conjoined through the merging of different assemblies, using the same mechanisms of ensemble formation. Sequence-order judgments appear to depend on prefrontal cortex (Petrides 1995). Sequence-order neurons are seen in medial premotor cortex (Merchant et al. 2013) and the hippocampus during spatial exploration and memory tasks (Kraus et al. 2013; Aronov et al. 2017).

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

Constructing Brackets

Brackets can be formed by different mechanisms. One possible mechanism is *attention* that selects one (out of many possible chunks) subset and then, through competition and winner-takes-all mechanisms, selects particular conjunctions over others. If these conjunctions contain a sequence, the network would automatically expect (produce) the sequence with the highest transition probabilities between successive states. This could, in part, address representing ordinal sequencing. However, this hypothesis is only pertinent when transition probability is a critical attribute of a sequence, and this is not always true, as in language.

An alternative solution is to distribute the bracketing task over different areas. We call this *anatomical factorization*, whereby the mapping rules for convergence determine the grain of the representations. Through convergence, the bracket around a chunk would now be a whole object. In this case, to get back to the relation of the components within the bracket, one would need to read out the nodes within the bracket, for instance through (top-down) feedback. To select the correct nodes at the lower level, some mechanism needs to be implemented that relates them to the big chunk in the bracket. This could be done by synchronizing ensembles across levels (as could be necessary, e.g., for mental imagery or silent speech).

Another possibility to form brackets is to use *time as coding space* and establish cross-frequency coupling. One could conceive slow rhythms as the bracket around a big chunk and the components to be represented by ensembles oscillating at higher frequencies. If consistent phase relations are assured between the slow and fast oscillations, it would be possible to decode which components belong to which of the bigger chunks within the bracket. Since recurrent networks can cope with the representation of sequences, the problem of ordinal coding can, in principle, be solved. Likewise, by having coupling across several different frequencies, nested relations can be specified. Such approaches to cross-frequency coupling (Hyafil et al. 2015), exemplified by the Lisman model (Lisman and Idiart 1995; Lisman 2005), are undoubtedly interesting, and perhaps even relevant, for some aspects of perception and cognition (e.g., Giraud and Poeppel 2012; Heusser et al. 2016). For this proposal, evidence is, however, scarce and contested.

Summary

The problem of language, along with the set of sequential operations that it entails, illustrates how investigating the way in which (sequential) relations are encoded in cortex and other domains may inform us about a uniquely human cognitive function. Successfully employing this strategy involves asking to what extent underlying operations are shared across domains (e.g., with spatial

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

navigation), and delineating which cognitive operations and neural codes are indeed unique to humans and which ones might be shared across species.

The Formal Basis of Generative Models for Language

In the previous section, we established some key aspects of the way in which language is encoded, and the unique compositional architecture these codes must possess. Here, we consider the form of *computations*, with a special focus on the cognitive operations entailed by language processing. Again, in the spirit of Marr (1982), it seems imperative to consider this level of description to guide our search for the neural implementation of cognitive functions. We will entertain the concept of a normative perspective (i.e., framing all computation under the overarching imperative to optimize, in some sense, the encoding of beliefs) as an alternative framework to help understand not only the computations subserving language but also any form of computation. This optimization is defined in terms of an objective function that has various interpretations in terms of information theory (i.e., self-information), self-organization, and self-evidencing.

Crucially, this optimization can be cast in terms of inference (namely, optimizing beliefs that are parameterized or encoded by neuronal quantities) and brings about the concept of a generative model, which is necessary to define the objective function. Still, the question remains: What sorts of generative models might be used by the brain to parse, synthesize, and generate language? We will focus on the distinction between generative models of continuous and discrete states and how they lead to different forms of optimization and message passing. This is an important distinction, as the type of generative models apt for language processing rests upon discrete states of the world, equipped with symbolic or semantic labels. Finally, we turn to the implications for cortical computation in terms of the message passing required for the ordinal and nested structures above. The structure of these models will turn out to be a key attribute that defines the challenges for understanding—and modeling—language processing in the brain. Aspects of this structure include the difficult problem of structure learning, the accommodation of structural dependency in linguistics, and the way we carve nature at the joints—via a nested factorization of the latent causes of language (and, in more general terms, any narrative that underlies our active engagement with the world "out there").

Encoding, Decoding, and the Neuronal Code

Modern versions of encoding (i.e., the mapping of a given stimulus onto a neural response) are associated with the notion of unconscious inference. On this view, there is a distinction between states of the world "out there" and the

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

sensory consequences of those states that are registered by sensory epithelia. Neuronal activity is taken to parameterize probabilistic beliefs over states of the world that are inferred through sensory impressions. If the neuronal code encodes beliefs about hidden or observed states of the world, what is computation?

One can develop a formal definition of neuronal computation in terms of an objective function that represents a lower bound on the evidence for a model and ensures the neuronal code is optimized with respect to states of affairs in the world. In Figure 14.2 (this volume), Karl Friston shows that this model is a probability distribution over the hidden states (i.e., causes) that generate sensory samples (i.e., consequences). Based on this model, one can compute the bound and specify neuronal dynamics in terms of a gradient descent of the ensuing objective function. This has a number of fundamental implications. First, it means that it should be possible, in principle, to specify neuronal dynamics in terms of self-evidencing (active inference) under some generative model. This implies that phenotyping a particular brain or creature boils down to specifying the generative model being used to navigate in their world. Under this view, emphasis is placed on understanding the form and nature of the generative model. Everything else should ideally follow from this model.

The second key observation is that any (probabilistic) generative model can be expressed as a Bayesian graph with nodes and edges. This is a simple construct that associates all hidden states (and sensations and actions) with nodes of a network, where the connections or edges denote conditional dependencies. The key point is that the form of the generative model defines, unambiguously, the requisite message passing among the nodes that constitute the gradient descent or neuronal dynamics. In other words, knowing the form of the generative mode means that we immediately know the computational or functional architecture of the brain, under the assumption that it is optimizing its beliefs about its world.

Thus, from basic principles we can arrive at a formal (if abstract) description of a brain that must be describable in terms of message passing among the nodes of a graph or network. Furthermore, in virtue of the causal structure in the world, the edges or connections will have a particular sparsity form (e.g., hierarchical structure). This means that we would expect to see self-evidencing computations play out on a relatively sparse (e.g., hierarchical) neuronal network. This resonates with the known neuroanatomy and neurophysiology of brains (Felleman and Van Essen 1991), which have this peculiar graphical structure, and could be contrasted with other organs such as the liver or blood. So what attributes might the generative model have?

At this point, we may consider the distinction between generative models of *discrete* and *continuous* states. This is a simple yet critical distinction based on the event space or support of the probability distributions (or densities). We can have states of the world that are categorical. In other words, we can

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

be in one room or another room, but not both rooms at once. Our beliefs then would be a categorical *distribution* over a finite set of states. Alternatively, states can be continuous. The analogous probability *density* over some continuous state (e.g., the luminance contrast at a particular point in the visual field) yields variables ranging from 0 to infinity. The corresponding probability density may have some (e.g., lognormal) distribution depending on the uncertainty about the actual level of luminance contrast. In terms of the neuronal code, this means that if we adopt discrete state space models, one might associate neuronal (population) activity with the probability of being in a particular state at any particular time. Conversely, in continuous state space models, neuronal activity may encode the expectation or average of the probability distribution and scale with the intensity or level of the continuous hidden state (e.g., luminance contrast). In both cases, the gradient descent to understand neuronal dynamics applies. However, the nature of these dynamics depends sensitively on whether our generative model is over discrete or continuous states.

Continuous state space models would call upon some form of Bayesian filtering to implement gradient descent when sensory input fluctuates over time. Common examples of these belief updating schemes include predictive coding, Kalman-Bucy filtering, particle filtering, unscented filtering, and their hierarchical (and nonlinear) variants. Analogous schemes for *discrete state space models* include belief propagation and variational message passing. Variational message passing corresponds to the solutions to the neuronal dynamics that explicitly optimize variational free energy. Crucially, these are not filtering or predictive coding schemes; although they share many computational aspects.

Perhaps the most important aspect is that all belief updating schemes entail reciprocal message passing over the edges of the Bayesian graph. This has a fundamental implication for cortical computation. It means that reciprocal neuronal connectivity must (either directly or indirectly) be in play, if the brain engages in Bayesian belief updating. This is a strong constraint on neuronal dynamics, which mandates recurrent connectivity and reciprocal message passing between any neurons or neuronal populations that constitute sufficient statistics of conditional or posterior beliefs. A popular example here can be found in predictive coding: in this particular message passing scheme, prediction errors are passed forward (e.g., in cortical hierarchies), while descending predictions are reciprocated in the other direction. Exactly the same reciprocal or recurrent exchange is found in belief propagation and variational message passing.

Under this view the fundamentals of computational architectures in the brain rest upon the following:

- Adopt a *constructivist* perspective on neuronal computations so that neuronal activity encodes beliefs about something; namely, states of the world that generate sensations.

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

- Specify the *structure of a generative model*, which specifies the graphical form; that is, network architecture of reciprocal message passing or Bayesian belief updating.
- Formulate this *message passing* in terms of differential equations to specify the precise architecture and form of neuronal dynamics.

How would this recipe for understanding cortical computation, in terms of belief updating, play out in the context of higher cognitive functions such as language?

Deep Generative Models for Language

We have tried to reduce the problem of understanding cortical computation to understanding the structure of generative models that explain how sensations are caused. We have posited that particular structures of generative models are necessary for language and detailed five structural aspects implicit to the generation of language, ranging from the ability to generate transitions among discrete states to the hierarchal nesting or parsing of tree structures. Furthermore, language has to deal with structural dependency, which could involve ordinal transposition and a particular form of parsing best understood in terms of hierarchical trees and their attendant decompositions.

From this arise two key implications for generative models that underlie neuronal dynamics in language processing. First, we are dealing with discrete state space models (e.g., hidden Markov models, Markov decision processes, hierarchal Dirichlet process models and their extensions), which immediately tells us that representations (i.e., expectations) about states of the world in the future (and past) are needed to support sequential transitions. Second, we need a hierarchical structure that allows for chunking and chaining within a particular (ordinal) temporal frame of reference. The requisite of deep temporal models brings with it some interesting functionality, along with some deep problems.

The capacity to represent sequences over time means that variational message passing builds, in effect, beliefs about the future (and the past). For example, reading the first word in a sentence already sets up a hypothesis space over all subsequent words, in virtue of message passing forward in time (and back again). This reciprocal message passing has, in part, a forward and backward aspect, in the sense that there is an explicit representation of the future. From a computational or cognitive perspective, it means that we have the capacity to hold in mind possible outcomes that are plausible given the sequential evidence sampled so far. Perhaps more interestingly, it also means that we can update our beliefs about initial experiences in the past. This provides an important opportunity to test hypotheses generated under these sorts of generative models, using prospective and retrospective inference, and to respond to unexpected evidence (i.e., violations at different levels of abstraction).

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

Simulating the variational message passing (under deep temporal models of this kind) exposes many issues related to the neurophysiological correlates of language processing. Perhaps the most interesting is the synchronous and asynchronous updating implied by discrete models. This necessarily involves the separation of timescales: a *fast* timescale for the optimization process itself and a *slower* timescale for sampling each new discrete sensory sample (e.g., through saccadic eye movements while reading or articulation of phonemes while talking). This discrete sampling of the world may progress at a theta frequency, while fast updating probably occurs with time constants associated with faster, for example, gamma frequencies (Melloni et al. 2009; Wang 2010; Giraud and Poeppel 2012). Furthermore, the hierarchal structure of these models necessarily entails a separation of temporal scales at different levels (e.g., delay period activity in the prefrontal cortex, in relation to fast dynamics lower in the auditory system). Empirically, this suggests a nesting of faster frequencies in slower frequencies, when belief updating is observed electrophysiologically with, necessarily, cross-frequency coupling and nested oscillations.

What are the special problems that accompany this sort of deep temporal model? These relate to the very structure or carving of this (linguistic) nature at its joints. The relational aspect of linguistic constructs (i.e., hidden causes or states) introduces a special problem that is probably best conceived of as a combinatorial explosion (e.g., discrete infinity). So what does this mean for the structure of the generative model?

One can finesse the complexity cost implicit in a combinatorial explosion by factorizing the generative model into conditionally independent causes and then binding these causes together, through convergent connectivity, to explain the particular pattern of sensory inputs at hand. A detour to vision may serve to clarify this point: an efficient way to address combinations of features (e.g., what and where) of an object in the visual field (i.e., low complexity encoding) would be to represent the nature (*what*) and location (*where*) attributes separately, then use the interaction or conjunction of these posterior expectations to predict the sensory input that would be sampled at any particular location in the visual field (Friston and Buzsáki 2016). This interaction between (roughly) orthogonal representational factors (i.e., a computational binding) entails second-order or multiplicative interactions between messages from the *what* and *where* parts of the generative model (e.g., the *what* and *where* pathways in the brain). In turn, this necessitates some form of modulatory or nonlinear optimization of synaptic efficacy of the sort associated with attentional selection mediated through dynamical mechanisms, as in communication through coherence (Fries 2005) or other neuromodulatory mechanisms. Thus, one important constraint of this view is that factorization implies multiplicative interactions when factorized features are combined.

What sort of factorization is in play in language? As indicated above, this factorization may be extremely complicated and must be hierarchically nested.

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

It might appear that certain syntactical structures are separated, by virtue of being associated with hidden factors from the actual semantic or phonological content. Furthermore, one has to consider the ordinal structure-dependent aspects of language. This speaks to the interesting possibility that we represent order or ordinal attributes in the same way that we represent locations in space. Put simply, there may be dedicated streams for encoding *when* that are combined with other factors encoding *what* at each level of hierarchal construction (Auksztulewicz et al. 2018). This is not unrelated to the notion of ordinal pointers involving convergent interactions between cortical language areas and the hippocampus (Friston and Buzsáki 2016).

At this point, one could start to speculate about the nested hierarchal and factorial form of generative models that would be fit for purpose in generating language. Perhaps, the most difficult problem in understanding the cortical computations that underlie language processing might not be in the details of the message passing or the biophysical implementation of the algorithms, but in understanding the very structure of the generative model and how this is acquired by a brain. This is known as structure learning or Bayesian model selection. These considerations emphasize the basic structure of generative models that possess the right sort of symmetry (i.e., invariances in conditional independencies) implicit in the right sort of carving or factorization. At present, simple symmetries have proven very effective in machine learning. Perhaps the most celebrated example of this is the weight sharing implicit in deep convolutional neuronal networks. This employs a simple factorization or invariance assumption that the weights of lateral connections at each level of the deep network are conditionally independent of their translational position. For the above arguments, we may be pressed to look for much more sophisticated symmetries that underlie our ability to parse and decompose invariance, when generating narratives in a world populated by creatures like us (who talk a lot). Let us now take a closer look at this issue from the perspective of neuroscience and AI.

Every Happy Marriage Has Its Ups and Downs: Neuroscience and AI

Apart from studying the brains of humans and other animals, a complementary inroad into understanding the computations that underpin human cognition may lie *in silico*. Comparatively recent advances in computer algorithms and hardware have led to a massive increase in the capacity of computers to fulfill tasks at human or even superhuman performance levels in domains such as the recognition of images, letters, or speech (LeCun et al. 2015) to playing computer games (Mnih et al. 2015) or even Go (Silver et al. 2016). At the forefront of these advances are deep neural networks (DNN); that is, hierarchical stacks of convolutional neural networks. Because their performance in certain domains is so close to or even better than that of humans and can be built at

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

will, DNNs seem to be promising tools to understand something more about human capacities. Especially in the domain of visual object recognition, DNNs now readily perform at the same level as humans or monkeys; interestingly, the properties of units in the higher layers of these networks show similar properties as neurons at the highest stages of object processing in monkey inferotemporal cortex (Yamins et al. 2014). This example suggests that there could be a fruitful, bidirectional exchange between AI and neurobiology to improve algorithms and to advance our understanding of the brain.

However, there are inherent differences between DNNs and human brains/behavior that are worth considering. For example, in terms of behavior: (a) deep learning algorithms need massive amounts of labeled training data (and regularization, etc.), whereas humans learn quickly and often in an unsupervised fashion; (b) DNNs typically learn specific tasks (e.g., recognizing cats) and generalize poorly to other, even similar tasks; (c) they do not have "common sense" and the domain of transfer learning is only emerging (Davis and Marcus 2015). In terms of biology, DNNs have many, sometimes hundreds of layers—more than the brain (e.g., the visual system is thought to consist of about 30 areas)—making it seem impossible to fit this number of layers into a skull. Furthermore, DNNs often rely on processing in massive data centers; running them on a small, autonomous device such as a phone immediately drains the battery. This serves to illustrate that brains and algorithms have evolved under different environmental pressures. Other long-standing arguments are that the backpropagation algorithms used to train DNNs are deemed biologically implausible (Crick 1989), although backpropagating action potentials and backward spread of plasticity have since been discovered (Fitzsimonds et al. 1997; Tao et al. 2000; Du et al. 2009). In addition, DNNs usually do not involve recurrent and long-range connections, which are characteristic of the cortex. Overall, this suggests that much remains to be accomplished before we can build machines that think and learn like humans (Lake et al. 2017). Still, comparing commonalities and differences between *in vivo* and *in silico* approaches to intelligence may be similarly fruitful as comparisons between species. What have new AI tools contributed to theories about human brain function? What do we learn about the brain by applying machine learning to neural data?

Much of AI today builds on neural networks models that were developed in the 1980s to understand features of human cognition based on aspects of what was known at the time about neural computations (e.g., multilayer structure, proximal connectivity). Hence, it may not be too surprising to find similarities between well-studied aspects of the neural processing and the way DNNs process data. One could thus argue that there is a fundamental circularity that we, neuroscientists, should remain aware of as we use these tools. These models were originally supposed to help us generate new testable hypotheses. Since these models were able to solve some (simple) computational operations, they have been reused for engineering purposes and refined to optimize machine

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

performance, no longer considering neurophysiological plausibility. This pragmatic and laudable use of neuronal networks or other brain-like algorithms may become problematic when we start applying them to "model" or "analyze" the brain. The constraints these models impose and the hypotheses they imply become implicit, and ignoring them might profoundly mislead us. For example, the apparent solution of using multiple layers to address data complexity may not be the (only) solution the brain uses to solve the same problem. Alternatively, using multiple layers may, at a certain level of abstraction, simply be a different formulation of the same solution the brain is also believed to use (Liao and Poggio 2016).

Although they are increasingly being used as analysis tools in neuroscience, what do machine learning techniques actually tell us about the brain? Research on animals and in humans with brain lesions has taught us that we need to know what is necessary as well as what is sufficient (Bouton et al. 2018). In animals, we get this information by considering both loss and gain in function studies. As this cannot be done in humans, computational approaches may help to address this question, building networks/models and perturbing them to try to find out what is necessary and sufficient for people to do a task. As discussed, the specific case of language is particularly challenging. To test whether some of the predictions regarding how humans process language are plausible, therefore, we have to build a model and then show that the predictions it makes, regarding neural responses to novel stimuli, are accurate. If successful, the model will have captured some of what actually happens in the human brain. In recent work, Pereira et al. (2018) developed a decoding model based on a limited amount of training data and showed that it can infer the meaning of new words, phrases, or sentences from patterns of brain activation. To do this, they described a high-dimensional semantic space and used a representative sample from this semantic space. If this is indeed how the brain represents such relationships, then a decoder trained in this way should be able to generalize from a relatively small training set to new concepts/relationships (as these are all dimensions of the semantic space). Pereira et al. were able to show that their decoder, trained only on a limited set of individual word meanings, can use this strategy to decode meanings of sentences in this way. These representations allow the decoder to distinguish between semantically similar sentences as well as to capture the similarity structure of inter-sentence semantic relationships. Thus, it may be a method by which the brain itself carries out these computations. This illustrates how such a method can be used to generate new hypotheses for neuroscientists to test in the actual brain, much along the lines for which these models were originally developed.

Another example of how machine learning paradigms could be used to tell us something about the brain lies in their ability to rescue function, again something that previously has mainly been shown in animals. For instance, Ezzyat et al. (2018) have shown that one can use a closed-loop system

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

to decode intracranially recorded neural activity from humans while they were learning lists of words, and then to implant artificially memories into lateral temporal cortex based on the patterns the machine learning algorithm extracted. Specifically, the system learned patterns associated with both successful encoding/recall and unsuccessful encoding/forgetting. Once the system had learned, the algorithm could then test whether or not this information was sufficient to induce memory: when it detected a pattern associated with forgetting, it stimulated the patients' brain to induce memory. Measured in terms of behavioral outcomes (i.e., words remembered), Ezzyat et al. showed that lateral temporal cortex—the site of stimulation—is sufficient to induce recall in humans.

Taken together, AI seems to offer both promises and pitfalls for neuroscience. Trying to understand what a DNN does comes with its own caveats. Clearly, neuroscientists should not naively apply DNNs for model building or analysis without considering the design principles of these networks. Nevertheless, reverse engineering neural networks that can solve tasks at human-level performance may provide a unique opportunity to grasp algorithmic and computational aspects of human intelligent behavior. As such, artificial neural networks should perhaps be treated like another species and not like a one-to-one model of the human brain. Finally, neuroscience should continue to build models that are solely made of a biological plausible set of submodels/routines, agnostic to neuroengineering tools, and provide biologically plausible options for engineering new algorithms.

Desiderata for the Future

We end our discussion by considering desiderata for future studies with a focus on pressing opportunities for further discoveries:

1. Understanding the coding of relations: The coding of relations between objects is a common theme across domains: vision, motor control, spatial navigation, cognitive/semantic maps, language, etc. How relations are coded on the fly for flexible and purposeful behavior remains, however, one of the next frontiers of knowledge. At the same time, whether similar or different mechanisms are repurposed for the encoding of relations across domains is unknown. Future studies will hopefully be able to close this gap in our understanding of a fundamental brain operation.
2. Development of mesoscopic measurements: Human neuroscience studies rest upon noninvasive, macroscopic measurements (e.g., fMRI, MEG, EEG) that are detached from the detailed microscopic measurements found in animal models. The wonderful assortment of (molecular) tools used in rodents and increasingly in nonhuman primates to understand mechanistically cortical circuitry and operations (and,

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

where possible, their causal relevance for behavior) cannot be used in humans, preempting a mechanistic understanding of the same processes and principles at the same level directly in the human brain. The development of measurement technologies at the mesoscopic scale that are safe and minimally invasive (e.g., multicontact recording arrays) may help bridge some gaps between human and animal studies and are needed more than ever. Progress in understanding the human brain may be fundamentally impeded without the development of such tools.

3. Ecological validity of behaviors in the laboratory versus in the wild: We question whether the experimental designs/tasks currently being used in much of neuroscience inappropriately constrain the type of answers that one might get. Specifically, how can we be certain that animals do or do not exhibit a specific behavior? Reductionist experimental paradigms, the reward schedules used to motivate animals to perform, as well as other variables may provide us with misguided answers simply because they do not tap into behaviors that an animal is equipped to produce. A possible alternative would be to access behaviors that intrinsically motivate animals to perform and exhibit specific behaviors. For example, we ask whether nonhuman primates would exhibit primitive forms of combinatorics when encouraged to teach conspecifics. Another venue for exploration would be to study behaviors in the wild, as those relate to the specific needs of the animals for survival. Inherently related to this question is whether training animals to perform human-like behaviors is informative or misleading our efforts to understand whether behavior across animals is similar or different.
4. Targeted and explicit interspecies comparisons: Darwin's idea that differences between species are a matter of gradation permeates most of the scientific practice. It is implicitly assumed that mechanisms will translate across species once we understand the evolutionary changes that have occurred. In practice, parallel strains of studies on rodents, nonhuman primates, and humans are often conducted without an adequate exchange or engagement between groups to permit explicit interspecies comparisons. Of course, such comparisons come with challenges. As discussed, homologies and analogies need to be carefully delineated using multimodal evidence; a focus on only one aspect of the organism (e.g., only brain structure) without considering other relevant factors (e.g., mechanics of the body, genetics, development) can be misleading. Especially when insights from preclinical animal studies (e.g., on psychiatric or neurological diseases) are to be translated to humans, evolutionary factors need to be more explicitly considered; this could rescue a human treatment doomed for failure because of a key evolutionary change that occurred after the split from a common ancestor to murine species. For many neuroscience questions, technologies that allow explicit interspecies comparisons are

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3

already available (e.g., comparative fMRI and intracranial recordings) and should be increasingly used for that purpose.

5. Illumination of canonical principles in minds and machines through AI: Perhaps we did not recognize how hard the problem of visual recognition was until we tried to build a machine that could do it. Thus, an attempt at building a machine (AI) that could exhibit comparable behaviors to those of the human brain may be a fruitful approach to understand basic principles of human cognition. Such a research program entails using computational models tested on increasingly exquisite sets of behaviors to decide, among the family of models, which model best approximates human behavior. Principles extracted from those models could be used as hypotheses for further cognitive experiments, to help guide additional insight into the computations performed. In parallel, efforts should be made to relate properties of the computational models explicitly to neural architecture, and the other way around (e.g., Nayebi et al. 2018). This is certainly not an easy task, and whether efforts will be successful remains to be determined.

These are exciting times in neuroscience. Over thirty years have passed since the seminal Dahlem Workshop on the neurobiology of neocortex (Rakic and Singer 1988). Although we are far from a full understanding of brain function and how it enables cognition, we are optimistic that the next thirty years will bring important insights. The right ingredients are there: a rapid pace in the development of neurotechnologies for studying the brain, a flourishing field in AI, the capacity to build algorithms that match human behavior, and a scientific community that is willing to rethink how cross-species comparisons are used to understand what the cerebral cortex does, how it evolved to do so, and how it can afford high-level cognition. Together, this holds promise in helping us understand how the cerebral cortex and its rich set of connections operates, and how this makes us human.

From "The Neocortex," edited by W. Singer, T. J. Sejnowski and P. Rakic.
Strüngmann Forum Reports, vol. 27, J. R. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-04324-3