Computational Cognitive Neuroscience

Carlos Zednik
carlos.zednik@ovgu.de
Otto-von-Guericke-Universität Magdeburg

forthcoming in M. Sprevak & M. Colombo (eds.),
The Routledge Handbook of the Computational Mind. London: Routledge.

1. Introduction

Computational cognitive neuroscience lies at the intersection of computational neuroscience, which aims to describe structures and processes in the brain through computational modeling and mathematical analysis, and cognitive neuroscience, which aims to explain behavior and cognition through the identification and description of neural mechanisms. Computational cognitive neuroscience invokes the descriptive tools of the former to achieve the explanatory aims of the latter: Computational models and mathematical analyses are used to identify and describe not just any structures and processes in the brain, but just those structures and processes that constitute the mechanisms of behavior and cognition.

Like investigators in other branches of neuroscience, computational cognitive neuroscientists rely on neuroscientific measurement techniques such as single-cell recording, functional magnetic resonance imaging (fMRI), and electroencephalography (EEG). Much more so than their colleagues in other branches of the discipline, however, computational cognitive neuroscientists additionally invoke formal methods developed in theoretical disciplines such as artificial intelligence, machine learning, statistics, mathematical physics, and the science of complex systems. These formal methods contribute to the aims of computational cognitive neuroscience in at least two ways. For one, they allow researchers to describe mechanisms not merely as consisting of certain neural structures and processes, but also as possessing particular computational, dynamical, and/or topological properties. For another, these formal methods facilitate the task of discovering such mechanisms in the first place. For example, if an algorithm is known to be particularly effective for simulating behavior and cognition on a computer, it may inspire computational cognitive neuroscientists to look for implementations of similar algorithms in the brain.

This chapter provides a methodological overview of computational cognitive neuroscience, centering on a distinction between two widely-used research strategies. On the one hand, *top-down* (or "reverse-engineering") strategies are used to infer, from formal characterizations of behavior and cognition, the function and structure of neural mechanisms. On the other hand, *bottom-up* strategies are used to identify and describe neural mechanisms and their formal properties, and to reconstruct their contributions to specific kinds of behavior and cognition. Although both research strategies simultaneously rely on neuroscientific measurement techniques and formal methods, they do so in markedly different ways. Moreover, both strategies can be used to understand cognitive systems at several different *levels of analysis* (Marr 1982), and to thereby deliver *mechanistic explanations* of these systems' behavioral and cognitive capacities (Bechtel 2008; Craver 2007; Zednik 2017).[1]

In what follows, the top-down and bottom-up research strategies will be contrasted through a series of examples. These examples also illustrate the diversity of formal methods being used, including methods to approximate Bayesian inference, methods to characterize stochastic processes, artificial neural network models, and analytic techniques from graph theory, dynamical systems theory, and information theory. Each example shows how computational cognitive neuroscientists go about discovering and describing the mechanisms responsible for specific behavioral and cognitive phenomena. At the same time, these examples reveal the characteristic limitations of the top-down and bottom-up strategies. Ultimately, explanatory success in computational cognitive neuroscience may in fact require a bidirectional approach.

2. Starting with behavior: Top-down strategies

One of the most widespread research strategies in computational cognitive neuroscience is a top-down (or "reverse-engineering") strategy inspired by David Marr's influential work on visual perception (Marr 1982). Marr sought to understand the visual system by analyzing it at three distinct *levels of analysis* (see also Chapter 15). At the *computational* level of analysis, he sought to answer questions about what the system is doing and why it is doing it. These questions are answered by specifying a mathematical function that describes the system's behavior, and by determining the extent to which this function reflects a relevant property or regularity in the environment (Shagrir 2010). At the *algorithmic* level, Marr considered questions about how the system does what it does. These questions can be answered by specifying the individual steps of an algorithm for computing or approximating the mathematical function that describes the cognitive system's behavior. Finally, at the *implementational* level of analysis, Marr was concerned with questions about where in the brain the relevant algorithms are actually realized, by identifying individual steps of an algorithm with the activity of particular neural structures. By analyzing the visual system at all three levels of analysis, Marr sought to simultaneously describe the physical and

[1] It will be assumed that (many) computational cognitive neuroscientists aim to deliver mechanistic explanations in the sense recently explored in the philosophy of neuroscience (Bechtel 2008; Craver 2007), and that the use of formal methods is in no way antithetical to this aim (see also Bechtel and Shagrir 2015; Piccinini and Craver 2011; Zednik 2017).

computational properties of the mechanism responsible for visual perception (see also Bechtel and Shagrir 2015; Piccinini and Craver 2011; Zednik 2017).

Although Marr deemed all three levels critical for the purposes of "completely understanding" a cognitive system (Marr 1982, 4), he argued that the best way to develop such an understanding would be to begin by answering questions at the computational level and to work downwards:

> "Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, which is critically important from an information-processing point of view. The reason for this is that [...] an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by understanding the mechanism (and the hardware) in which it is embodied." (Marr 1982, 27; see also Dennett 1994)

Thus, Marr's top-down strategy involves using answers already available at higher levels of analysis to constrain the answers that have might be given to questions at lower levels (Zednik and Jäkel 2016). In other words, reverse-engineering is a matter of *inferring* the function and structure of mechanisms from (among others) prior characterizations of the behavioral and cognitive phenomena for which they are deemed responsible.[2]

Many past and present research efforts in computational cognitive neuroscience pursue this kind of reverse-engineering strategy. To this end, they often take as their starting point characterizations of behavior and cognition previously developed in disciplines such as cognitive psychology and psychophysics. Frequently, these characterizations take the form of statistical models of behavioral data. For example, the investigation of perceptual decision-making introduced below aims to uncover the neural mechanisms responsible for the characteristic shape of response-time distributions in human subjects (Cao et al. 2016). Similarly, the studies of human categorization reviewed later in this section begin with a model in which the explanandum phenomenon is characterized as a form of Bayesian probabilistic inference (Anderson 1991b; Sanborn, Griffiths, and Navarro 2010). That said, computational-level characterizations of behavior and cognition need not be statistical; many reverse-engineers in computational cognitive neuroscience begin with characterizations of behavior and cognition as forms of information-processing, in which inputs are deterministically transformed into outputs (e.g. Marr and Hildreth 1980), or as dynamical trajectories through a multidimensional state-space with characteristic regions of stability and instability.

[2] This top-down inference need not be completely unconstrained by low-level considerations, of course. Indeed, Marr himself often appealed to extant knowledge of neurological structures in addition to computational-level considerations. Nevertheless, as Marr's own words illustrate, it is characteristic of the top-down approach that the latter be weighted more heavily than the former.

Investigators often have a choice to make about how to describe an explanandum phenomenon at the computational level. For example, an agent's reaching behavior might be characterized probabilistically, as the likelihood of reaching toward a particular target, but also dynamically, as a continuous trajectory through space and time. Such descriptive choices are not without consequence; the particular way in which a phenomenon is described can have a profound effect on the mechanisms that are likely to be discovered. This is because, given a formal characterization of the explanandum phenomenon at the computational level, the top-down strategy proceeds by identifying one or more algorithms with which to compute or approximate the mathematical function used in that characterization. Algorithms may be viewed as descriptions of the functional processes that contribute to a cognitive system's behavior: The component operations and functional organization of the mechanism responsible for that behavior (Zednik 2017). Unfortunately, the identification of algorithms is often hampered by a considerable degree of uncertainty: Many different algorithms serve to compute or approximate any particular mathematical function, and investigators are tasked with identifying the algorithm that is *actually* used by the relevant cognitive system, from among many possible algorithms that it might *possibly* use (Piccinini and Craver 2011). In order to deal with this kind of uncertainty, many advocates of the top-down approach deploy *heuristics* to constrain the search space of possible alternatives (Simon 1998; Zednik and Jäkel 2016). Although fallible—the chosen heuristic might highlight an algorithm that is not actually implemented by the target system—these heuristics are instrumental for the purposes of efficiently formulating testable hypotheses at the algorithmic level of analysis.

One intuitive heuristic for formulating testable hypotheses at the algorithmic level is the *mirroring heuristic*. This heuristic is deployed whenever investigators assume that functional processes in the brain exhibit the same mathematical structure as the explanandum phenomenon under a particular formal characterization. The use of this heuristic ensures that the particular mathematical formalism that is invoked at the computational level has a direct influence on the hypotheses that will actually be considered at the algorithmic level. Perhaps the clearest recent example of the mirroring heuristic at work can be observed in recent efforts to motivate the *Bayesian coding hypothesis* (Knill and Pouget 2004; Ma et al. 2006). Motivated by characterizations of behavior and cognition as forms of optimal probabilistic inference—in which sensory evidence is combined with prior beliefs in accordance with Bayes' rule (Anderson 1991a; Oaksford and Chater 2001)—proponents of this hypothesis argue that neural mechanisms themselves implement probability distributions, and compute over them using (close approximations of) Bayes' rule (see also Colombo and Hartmann 2015; Zednik and Jäkel 2014).

Although the mirroring heuristic may be intuitive and easy to deploy, it is also potentially misleading. As has already been stated above, many different algorithms serve to compute or approximate any particular mathematical function. Thus, there is no reason to believe, from behavioral evidence alone, that the brain actually implements just the one algorithm that most accurately reflects the mathematical structure of the phenomenon being explained (see also Maloney and Mamassian 2009). Moreover, in some cases the mathematical characterizations used at the computational level are such that the mirroring heuristic would yield algorithms that are psychologically or biologically implausible. For example, it is well-known that the generic algorithm for solving problems of optimal probabilistic inference via Bayes' rule is, in general, computationally intractable (Kwisthout, Wareham, and van Rooij 2011). For this reason, the explanatory success of the top-down strategy is likely to depend on the use of heuristics more nuanced than mirroring.

One such heuristic may be the *tools-to-theories heuristic* (Gigerenzer 1991). This heuristic is deployed whenever investigators assume that the algorithms implemented in the brain resemble an instrument, tool, or analytic technique that has previously been used to measure, study, or describe the behavioral or cognitive phenomenon being investigated. Notably, researchers in theoretical disciplines such as computer science, artificial intelligence, machine learning, and statistics have over time compiled a large portfolio of algorithms with which to compute or approximate many different mathematical functions in particularly efficient and/or reliable ways. Insofar as some of these functions resemble the ones that have been used to characterize behavior and cognition at the computational level, the tools-to-theories heuristic allows computational cognitive neuroscientists to exploit this portfolio for explanatory gains. As Griffiths et al. have remarked, "the best algorithms for approximating probabilistic inference in computer science and statistics" may be used as "candidate models of cognitive and neural processes" (Griffiths, Vul, and Sanborn 2012, 264).

Consider a recent example, also from the recently-prominent Bayesian approach. Sanborn et al. (2010) advance the hypothesis that the mechanisms for categorization as described by Anderson (1991b) implement a *particle filtering* algorithm—a kind of *Monte Carlo sampling* that has been developed in machine learning to approximate optimal Bayesian inference. To this end, Sanborn et al. evaluate the performance of this algorithm relative to two alternatives: *Gibbs sampling* and Anderson's own *iterative algorithm*. Like particle filtering, these alternatives are also co-opted from applications in machine learning and artificial intelligence. Unlike particle filtering, however, Sanborn et al. demonstrate that these alternatives do not produce the kinds of order effects that are typically observed in human behavior. Therefore, they postulate that the particle-filtering algorithm is more likely than the two alternatives to correctly describe the operations of the mechanism for human categorization.

The tools-to-theories heuristic has also been used within a broadly dynamical approach. In a recent study on bistable perception, Cao et al. (2016) evaluate the relative merit of four different stochastic processes for explaining the characteristic "reversals"—

spontaneous changes in the percept—that occur when human subjects encounter ambiguous stimuli such as the Necker cube. Each one of the *Poisson*, *Wiener*, *Ornstein-Uhlenbeck* and *generalized Ehrenfest* processes (for a review see Cox and Miller 1977) are mathematical models previously used in disciplines such as statistical mechanics and telecommunications to predict e.g. the emission of particles from a radioactive source and the arrival of calls at a telephone exchange. In computational cognitive neuroscience, Cao et al. show that a generalized Ehrenfest process, unlike the others, reproduces the kind of short-tailed and scale-invariant distribution of reversals that is typically observed in human behavior. Thus, by starting from a detailed characterization of the relevant behavioral dynamics, and evaluating the relative ability of four well-understood stochastic processes to reproduce these dynamics, Cao et al. invoke the tools-to-theories heuristic to advance a testable algorithmic-level hypothesis, viz. that the neural structures involved in bistable perception implement a generalized Ehrenfest process.

In a reverse-engineering context, the mirroring and tools-to-theories heuristics are used to descend from the computational to the algorithmic level of analysis. But given a particular algorithm, investigators still face a question about how that algorithm is implemented in the brain. Answering this question is a matter of identifying the steps of the algorithm with the activity of specific neural structures in the brain, so as to answer a question about where in the brain the relevant functional processes are carried out (Bechtel and Richardson 1993; Zednik 2017). Sometimes, this kind of identification proceeds quite directly, by invoking neuroscientific measurement techniques such as single-cell recordings or fMRI to identify neural structures that exhibit patterns of activity that can be correlated with the ones posited by the algorithm. For example, in one particularly influential study of perceptual decision-making, Newsome et al. (1989) show that psychophysical judgments by macaque monkeys in a random-dot motion-detection task are well-correlated with concurrent single-cell recordings in area MT, and for this reason conclude that single MT-neurons themselves perform a kind of signal detection. More recently, proponents of the Bayesian coding hypothesis have sought to identify the location of probabilistic representations in the brain via fMRI imaging (see e.g. Vilares et al. 2012).

In many other cases, however, answering a question at the implementational level involves a considerable degree of speculation. Indeed, making an educated guess about which structure in the brain *might* implement a particular algorithm is perhaps the most common way in which proponents of the top-down strategy formulate testable hypotheses about where in the brain a particular process is carried out. Consider, for example, Marr & Hildreth's (1980) discussion of visual edge-detection, in which they speculate how the detection of "zero-crossings" might be implemented in area LGN:

> "if an on-centre geniculate cell is active at location P and an off-centre cell is active at nearby location Q, then then the value of $\Delta^2 G * I$ passes through zero between P

> and Q. Hence, by combining the signals from P and Q through a logical AND-operation, one can construct an operator for detecting when a zero-crossing segment (at some unknown orientation) passes between P and Q. By adding nonlinear AND-operations in the longitudinal direction, one can, in a similar way, construct an operator that detects oriented zero-crossing segments." (Marr and Hildreth 1980, 208–9)

Notably, Marr & Hildreth's appeal to AND-operations being implemented by geniculate cells is entirely speculative, being motivated by considerations of how the brain *might* detect zero-crossings rather than by actual knowledge of LGN. In a similarly speculative way, Pouget et al. (2013) outline several different ways in which the brain *might* represent probability distributions so as to underwrite the Bayesian coding hypothesis: The firing rate of a single neuron could directly code log-probabilities; a population of neurons with differing tuning curves may code a probability distribution by a basis function expansion; or the activity of pools of neurons might represent samples from a distribution. Finally, in the context of bistable perception, Cao et al. (2016) suggest that the neural units most suited for implementing a generalized Ehrenfest process may be those that are assembled into so-called *attractor networks*, which are known to exist, but whose actual contribution to behavior and cognition remains unclear (Amit 1995).

These examples show that, whereas proponents of the top-down approach in computational cognitive neuroscience have recourse to a wide array of algorithms with which to compute a particular function, they tend to be quite limited in their knowledge of how these algorithms are actually implemented in the brain. On the one hand, this observation gives credence to Marr's original suggestion that it is often easier to model an algorithm by considering what cognitive systems do, than by reflecting on the neural structures in which those algorithms are likely to be implemented. Indeed, as statistics, computer science, artificial intelligence, machine learning, and other theoretical disciplines provide an increasingly detailed understanding of the relative efficiency and degree of optimality of different algorithms, it seems likely that these disciplines' influence on the course of neuroscientific research will continue to grow. On the other hand, the examples reviewed here also show why the top-down approach may ultimately prove unsatisfying: Although it has become relatively easy to formulate algorithmic-level hypotheses for a wide variety of phenomena, it remains difficult to know which of these hypotheses are actually true. It is in order to avoid this difficulty that, rather than begin with behavior and work their way down, many computational cognitive neuroscientists instead begin with the brain and work their way up.

3. Starting with the brain: Bottom-up strategies

When Marr professed the benefits of the reverse-engineering approach, he could not have predicted the degree to which technological advances would eventually transform the bottom-up strategy into a viable alternative. Rather than infer the function and structure of neural mechanisms from characterizations of the phenomena being explained, bottom-up strategies in computational cognitive neuroscience aim to explain these phenomena by reproducing them in models and simulations that incorporate functional and structural details from several levels of brain organization.[3] As such, these strategies rely on single-cell-recording, fMRI imaging, EEG and other neuroscientific measurement techniques that provide insight into the behavior, composition and organization of mechanisms at the level of individual neurons, neural populations, and/or cortical columns and regions. Moreover, they invoke computational modeling methods and methods of mathematical analysis to illuminate the relevant mechanisms' statistical, dynamical, topological, and/or computational properties. Insofar as these techniques and methods can be used to discover and describe mechanisms, and to show how these mechanisms give rise to specific behavioral and cognitive phenomena, they yield mechanistic explanations of these phenomena (Bechtel 2008; Craver 2007).

Because the bottom-up strategy is driven by the insights provided by neuroscientific measurement techniques, this strategy tends to be most effective when the relevant techniques are most reliable. Among the most reliable measurement techniques is the single-cell recording, a measure of electrical activity at the level of individual nerve cells. At least since the 1950s, neuroscientists have appealed to the activity of single cells to explain the behavioral and cognitive capacities of whole organisms. This approach has been particularly influential in the domain of visual perception, in which *feature detector* cells have been discovered whose activity correlates with specific environmental features such as moving edges (Lettvin et al. 1959), oriented bars (Hubel and Wiesel 1959), and (famously) bugs (Barlow 1953). Motivated by these results, Horace Barlow advanced a *single neuron doctrine,* according to which the computational capacities of individual nerve cells suffice to explain an organism's perceptual abilities: “The subtlety and sensitivity of perception results from the mechanisms determining when a single cell becomes active, rather than from complex combinatorial rules of usage of nerve cells” (Barlow 1972, 371).

Barlow's doctrine resonates even today. In an example briefly introduced above, Newsome et al. (1989) have found that recordings of individual MT neurons predict the performance of macaque monkeys in a random-dot motion-detection task. Motivated by this finding, the authors hypothesize that individual MT neurons solve the very same kind of

[3] Levels of organization should not be confused with levels of analysis. Whereas the former are individuated by the kinds of questions an investigator might ask about a particular cognitive system, the latter are individuated by constitution-relations within a mechanism (Bechtel 2008; Craver 2007). Insofar as many mechanisms are hierarchical it is often profitable to apply each one of the three levels of analysis at any single level of organization.

signal-detection task that is solved by the monkey as a whole. Although it may be questioned whether correlations between the activity of single cells and the behavior of whole organisms are really all that significant (Stüttgen, Schwarz, and Jäkel 2011), such correlations are still frequently appealed to in the context of the bottom-up approach in computational cognitive neuroscience: Investigators attempt to explain the behavior of whole organisms by showing how that behavior can be reproduced by mechanisms at the level of individual neurons.

That said, it is fair to question whether the computational capacities of individual neurons suffice to explain behavioral and cognitive phenomena in general. Indeed, it is now a commonplace to assume that performance in a wide variety of tasks—especially tasks further removed from the sensorimotor periphery such as planning, reasoning, language learning, and attention—requires the computational capacities of neural networks (Yuste 2015). Neural networks took center stage in computational cognitive neuroscience with the development of sophisticated *connectionist* modeling methods in cognitive science, in which networks of artificial "neural" units, arranged in layers and interconnected with weighted "synaptic" connections, are used to replicate various behavioral and cognitive capacities (see also Chapters 5 and 8). Of course, early connectionists stressed the fact that their models were highly idealized, and that they should for this reason be considered "neurally inspired" rather than biologically plausible (Thorpe and Imbert 1989). Nevertheless, many computational cognitive neuroscientists today rely on connectionist models that incorporate an ever-increasing degree of biological realism[4], thus allowing them to view these models as plausible descriptions of the mechanisms responsible for specific behavioral and cognitive phenomena.

Consider a recent attempt to explain *C. elegans klinotaxis,* a form of goal-directed locomotion in which the nematode worm approaches a chemical source by way of a regular oscillatory motion. Beginning with a complete description of the *connectome*—a graphical representation of the *C. elegans* nervous system at the level of individual neurons (White et al. 1986)—Izquierdo & Beer (2013) derive a *minimal network* that includes only those chemosensory, motor, and inter-neurons that, due to graph-theoretical considerations, are deemed most likely to contribute to the production of klinotaxis. By inserting the minimal network into a simulated *C. elegans* body model, and in turn situating that body model within a simulated environment (see also Izquierdo and Lockery 2010), Izquierdo & Beer artificially evolve network parameters suitable for the production of reliable and efficient klinotaxis. By comparing the klinotaxis produced in simulation to the klinotaxis produced in

[4] Many, but not all. Motivated in no small part by the finding that connectionist models with highly idealized and simplified "neural" units and connections are universal function approximators (Hornik 1991), these models have become widespread in engineering disciplines such as artificial intelligence and machine learning (see e.g. Schmidhuber 2014). Investigators working in these disciplines traditionally value computing power, efficiency, and optimality over biological realism.

the real world, Izquierdo & Beer advance the hypothesis that the minimal network is "appropriate for the generation of testable predictions concerning how the biological network functions" (Izquierdo and Beer 2013, 5). Indeed, in a subsequent study, Izquierdo, Williams and Beer (2015), propose that some of the interneurons in the minimal network constitute "informational gates" through which chemosensory information is allowed to flow and thereby influence motor neuron activity, but only at specific moments in time. This "informational gating" is postulated to be a crucial feature of the mechanism for the oscillatory nature of *C. elegans* klinotaxis not only in simulation, but also in the real world (for discussion see also Zednik, in press).

Of course, it is unclear whether Izquierdo and colleagues' approach will eventually scale up; investigators are still a long way away from having a comparable model of the *human* connectome (but cf. Sporns 2012). Nevertheless, computational cognitive neuroscientists have made great progress in adding biological detail to many different connectionist models. For example, rather than deploy networks whose units exhibit a sigmoidal activation function, many investigators today deploy networks of spiking units which exhibit time-varying response profiles reminiscent of biological neurons (Maass 1997). Moreover, many others deploy networks whose weights are determined by learning algorithms more biologically plausible than the backprogation algorithm developed in the 1980s (see also Chapter 5). Finally, reminiscent of the aforementioned work on *C. elegans* klinotaxis, some investigators no longer model generic neural network mechanisms, but rather aim to describe specific networks in well-defined areas of the brain (e.g. the hippocampus: Gluck and Myers 1997). In general, insofar as connectionist models can be used to reproduce specific behavioral or cognitive capacities while incorporating an ever-increasing degree of biological realism, they deliver plausible mechanistic explanations of these capacities.

Connectionist models are traditionally viewed as describing networks of interconnected neurons. A different family of network models aims to describe networks of interconnected columns and regions, distributed across the brain as a whole (Sporns 2011). Like other bottom-up approaches in computational cognitive neuroscience, the development of network models of this kind is grounded in knowledge of biological reality. Unfortunately, at this high level of brain organization, there is considerable disagreement about the reliability and informativeness of the measurement techniques that are used to acquire such knowledge. For example, there is still no agreed-upon method of individuating brain regions; investigators rely on a variety of *parcellation schemes* with which to identify individual network elements. Whereas some of these schemes may be quite principled—as when network elements are identified with Brodmann areas, themselves individuated on the basis of cytoarchitectural principles—other schemes are quite pragmatic, such as when network elements are identified with the location of electrodes in EEG recordings or with voxels in fMRI data. Similarly, investigators also rely on a variety of *connectivity schemes* for determining the extent to which any two network elements are connected. Whereas the elements of *structural* networks are connected anatomically, the elements of so-called *functional* networks have activity that is connected statistically, i.e. that is correlated over time. Most intriguingly, perhaps, the connections of *effective* networks correspond to the presumed causal interactions between network elements (Friston 2011), often operationalized in terms of information-theoretic measures such as *Granger causality*. Not surprisingly, the use of such a wide variety of parcellation and connectivity schemes has led to the proliferation of whole-brain network models, with little certainty about how these models actually contribute toward specific explanatory goals (Craver 2016; Miłkowski 2016; Zednik, in press). Thus, although bottom-up strategies at the level of the brain as a whole are grounded in an abundance of neuroscientific measurement data, it remains unclear to what extent this data constitutes genuine knowledge of the mechanisms responsible for behavior and cognition.

Although it remains unclear how this epistemological difficulty can be overcome, it is nevertheless worth understanding the way neuroscientific measurement data at the level of the brain as a whole can be analyzed using sophisticated mathematical and computational methods. These methods illuminate a particular network's topological, dynamical, and/or informational properties, and may also reveal potential interdependencies between different kinds of properties. In particular, graphs are frequently used to model a brain network's topology, and graph-theoretic techniques are used to identify the presence of e.g. hub nodes, modules, and motifs (Bullmore and Sporns 2009). Moreover, the results of graph-theoretic analyses are increasingly deployed to constrain network models not unlike the connectionist models reviewed above. Although the units in these network models correspond to cortical columns or regions (or pragmatically-individuated fMRI voxels) rather than to individual neurons, they can similarly be used to simulate the relevant network's behavior, and to compare that behavior to the properties of a phenomenon of explanatory interest. This kind of comparison is greatly facilitated by information-theoretic measures that illuminate e.g. the flow of information through a network (Izquierdo, Williams, and Beer 2015), as well as by dynamical systems theoretic techniques that characterize e.g. patterns of rhythmic oscillation, stable states, and bifurcations in the activity of individual network elements and/or in the behavior of the network as a whole (Uhlhaas et al. 2009). Perhaps the most interesting studies of this kind combine the insights of several different analytic techniques, thereby revealing dependencies between e.g. a network's topological properties and its behavioral dynamics (e.g. Pérez et al. 2011), or between its dynamical and informational properties (Beer and Williams 2015). Provided that computational cognitive neuroscientists are eventually able to overcome the epistemological difficulties associated with the identification of large-scale network mechanisms, these analyses of networks' formal properties are likely to deliver a detailed understanding of the way a network mechanism's behavior depends on its composition and organization. That is, they are poised to *show how* such mechanisms give rise to specific behavioral and cognitive phenomena (Craver 2016; Zednik 2014).

In general, these examples show that, no matter the level of brain organization, bottom-up strategies in computational cognitive neuroscience depend on the reliability and informativeness of neuroscientific measurement techniques, as well as on the descriptive power of computational modeling methods and methods of mathematical analysis. *Pace* Marr's concerns about the viability of the bottom-up approach, these techniques and methods render the bottom-up approach increasingly useful for uncovering the mechanisms for behavior and cognition. Indeed, insofar as they are grounded in knowledge of biological reality, bottom-up strategies are likely to be far more constrained than the top-down strategies discussed above. At the same time, bottom-up strategies have at least two characteristic limitations of their own.

For one, although bottom-up strategies are often more constrained than top-down alternatives, the descriptive power of the relevant mathematical and computational models still frequently exceeds the available knowledge of biological reality. This problem was widely acknowledged in the early days of the connectionist research program, but seems to have been mostly overcome due to the newfound ability to incorporate greater biological detail. That said, the same problem has once again come to the fore at the level of the brain as a whole. As illustrated by the lack of consensus about how to individuate the elements and connections of whole-brain networks, it has become relatively easy to identify and represent networks in the brain, but comparatively difficult to know which (aspects) of these networks should actually be cited in explanations of specific behavioral and cognitive phenomena. Indeed, commentators sometimes question the explanatory import of structural network modeling initiatives to map the *C. elegans* connectome, and Craver (2016) has recently denied that functional networks of pragmatically-individuated fMRI voxels should be viewed as explanations at all. Although these outright dismissals seem exaggerated—structural and functional network models might represent certain *aspects* of a mechanism, even if they do not represent the mechanism as a whole (Hochstein 2016; Zednik, in press)—bottom-up approaches in computational cognitive neuroscience are likely to properly get off the ground only when they are rooted in reliable knowledge of neurobiological reality.

For another, our ability to uncover neurobiological detail, and our ability to model that detail on a computer, may also often outstrip our ability to understand the mechanisms whose details are being modeled. Especially at the level of the brain as a whole, it is possible that computational models of network mechanisms remain *opaque* (Dudai and Evers 2014). That is, these models may be no more easily analyzed and understood than the relevant mechanisms themselves. Although it remains unclear to what extent a model's intelligibility is related to its capacity to explain, this harks back to Marr's original suggestion that it may be easier to identify the computational workings of the brain by considering what it does, than by describing what it is made of. Of course, as the preceding examples show, bottom-up approaches do not only rely on computational models, but also invoke sophisticated mathematical techniques to analyze the dynamical, topological, and informational properties of the mechanisms being modeled. It remains to be seen whether these techniques are sufficiently illuminating to reveal the inner workings of even the most complex and large-scale brain mechanisms (see also Zednik 2015).

4. Conclusion: Toward a bidirectional approach?

Many research efforts in computational cognitive neuroscience can be viewed as instances of either the top-down or the bottom-up research strategy. Top-down (or reverse-engineering) strategies aim to infer the function and structure of neural mechanisms from prior descriptions of behavior and cognition. In contrast, bottom-up strategies seek to reproduce behavioral and cognitive phenomena in computational models that are grounded in knowledge of biological reality. Both strategies have distinct advantages, but also characteristic limitations. Whereas top-down strategies have recourse to a plethora of mathematical formalisms and computational algorithms co-opted from disciplines such as artificial intelligence, machine learning, and statistics, they still regularly bottom out in speculative proposals about how such algorithms might actually be implemented in the brain. In contrast, because bottom-up strategies take as their starting point actual neuroscientific measurement data, they are not similarly limited by this kind of speculation. Nevertheless, these strategies are often limited by insufficiently informative empirical data —especially at the level of the brain as a whole—and by computational models that may be no easier to understand than the mechanisms they are supposed to be models of.

In closing, it is worth considering the possibility that the characteristic limitations of the top-down and bottom-up strategies might eventually be overcome by adopting something akin to a *bidirectional* approach. Indeed, practicing scientists are not beholden to conceptual distinctions, and are free to adopt aspects of both research strategies simultaneously. In fact, many investigators do so already. For example, when proponents of reverse-engineering speculate about the possible neural implementations of a particular algorithm, they do not do so in a vacuum, but actually rely on what they already know about neurobiological reality to constrain their own speculative proposals. As the available knowledge of neural mechanisms increases, the frequency of unconstrained speculation in the context of the top-down approach decreases.

In a similar way, top-down considerations may enable researchers to overcome some of the characteristic limitations of the bottom-up approach. Theoretical disciplines such as artificial intelligence, machine learning and statistics have not only developed a large portfolio of algorithms to be used as testable hypotheses, but have also developed sophisticated methods of mathematical analysis to understand how such algorithms actually work. For example, the popularity of *deep learning networks* in machine learning (Schmidhuber 2014) has led to the development of analytic tools that can be used to understand different levels of representational abstraction in hierarchical networks (e.g. Montavon, Braun, and Müller 2011). Although deep learning networks are generally considered biologically implausible because they, like early connectionist models, rely on backpropagation learning, it may be that some of the tools originally developed to understand the workings of deep learning networks can be co-opted to understand the computational capacities of hierarchical networks in the biological brain. In this way, a standard trick from the reverse-engineering toolbox—co-opting developments in theoretical disciplines such as machine learning—may even allow proponents of the bottom-up approach to overcome characteristic limitations such as opacity. More generally, therefore, it may be that the most fruitful research strategy for explanatory success in computational cognitive neuroscience is a bidirectional one.

5. References

Amit, Daniel J. 1995. “The Hebbian Paradigm Reintegrated: Local Reverberations as Internal Representations.” *Behavioral and Brain Sciences* 18 (4): 631–631.

Anderson, John R. 1991a. “Is Human Cognition Adaptive?” *Behavioral and Brain Sciences* 14: 471–517.

———. 1991b. “The Adaptive Nature of Human Categorization.” *Psychological Review* 98 (3): 409.

Barlow, Horace B. 1953. “Summation and Inhibition in the Frog’s Retina.” *The Journal of Physiology* 119 (1): 69–88.

———. 1972. “Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology?” *Perception* 1 (4): 371–394.

Bechtel, William. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Routledge.

Bechtel, William, and Robert C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. MIT Press ed. Cambridge, Mass: MIT Press.

Bechtel, William, and Oron Shagrir. 2015. “The Non-Redundant Contributions of Marr’s Three Levels of Analysis for Explaining Information-Processing Mechanisms.” *Topics in Cognitive Science* 7 (2): 312–22. https://doi.org/10.1111/tops.12141.

Beer, Randall D., and Paul L. Williams. 2015. “Information Processing and Dynamics in Minimally Cognitive Agents.” *Cognitive Science* 39 (1): 1–38. https://doi.org/10.1111/cogs.12142.

Bullmore, Ed, and Olaf Sporns. 2009. “Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems.” *Nature Reviews Neuroscience* 10 (3): 186–98. https://doi.org/10.1038/nrn2575.

Cao, R., A. Pastukhov, M. Mattia, and J. Braun. 2016. “Collective Activity of Many Bistable Assemblies Reproduces Characteristic Dynamics of Multistable Perception.” *Journal of Neuroscience* 36 (26): 6957–72. https://doi.org/10.1523/JNEUROSCI.4626-15.2016.

Colombo, Matteo, and Stephan Hartmann. 2015. “Bayesian Cognitive Science, Unification, and Explanation.” *The British Journal for the Philosophy of Science*, axv036.

Cox, D.R., and H.D. Miller. 1977. *The Theory of Stochastic Processes*. New York: CRC Press.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

———. 2016. “The Explanatory Power of Network Models.” *Philosophy of Science* 83 (5): 698–709.

Dennett, Daniel C. 1994. “Cognitive Science as Reverse Engineering: Several Meanings of ‘Top-Down’ and ‘Bottom-Up.’” *Logic, Methodology and Philosophy of Science* IX: 679–89.

Dudai, Yadin, and Kathinka Evers. 2014. “To Simulate or Not to Simulate: What Are the Questions?” *Neuron* 84 (2): 254–61. https://doi.org/10.1016/j.neuron.2014.09.031.

Friston, Karl J. 2011. “Functional and Effective Connectivity: A Review.” *Brain Connectivity* 1 (1): 13–36. https://doi.org/10.1089/brain.2011.0008.

Gigerenzer, Gerd. 1991. “From Tools to Theories: A Heuristic of Discovery in Cognitive Psychology.” *Psychological Review* 98 (2): 254.

Gluck, Mark A., and Catherine E. Myers. 1997. “Psychobiological Models of Hippocampal Function in Learning and Memory.” *Annual Review of Psychology* 48 (1): 481–514.

Griffiths, Thomas L., Edward Vul, and Adam N. Sanborn. 2012. “Bridging Levels of Analysis for Probabilistic Models of Cognition.” *Current Directions in Psychological Science* 21 (4): 263–68. https://doi.org/10.1177/0963721412447619.

Hochstein, Eric. 2016. “One Mechanism, Many Models: A Distributed Theory of Mechanistic Explanation.” *Synthese* 193 (5): 1387–1407. https://doi.org/10.1007/s11229-015-0844-8.

Hornik, Kurt. 1991. “Approximation Capabilities of Multilayer Feedforward Networks.” *Neural Networks* 4: 251–57.

Hubel, David H., and Torsten N. Wiesel. 1959. “Receptive Fields of Single Neurones in the Cat’s Striate Cortex.” *The Journal of Physiology* 148 (3): 574–591.

Izquierdo, Eduardo J., and Randall D. Beer. 2013. “Connecting a Connectome to Behavior: An Ensemble of Neuroanatomical Models of C. Elegans Klinotaxis.” Edited by Lyle J. Graham. *PLoS Computational Biology* 9 (2): e1002890. https://doi.org/10.1371/journal.pcbi.1002890.

Izquierdo, Eduardo J., and S. R. Lockery. 2010. “Evolution and Analysis of Minimal Neural Circuits for Klinotaxis in Caenorhabditis Elegans.” *Journal of Neuroscience* 30 (39): 12908–17. https://doi.org/10.1523/JNEUROSCI.2606-10.2010.

Izquierdo, Eduardo J., Paul L. Williams, and Randall D. Beer. 2015. “Information Flow through a Model of the C. Elegans Klinotaxis Circuit.” Edited by Gennady Cymbalyuk. *PLoS One* 10 (10): e0140397. https://doi.org/10.1371/journal.pone.0140397.

Knill, David C., and Alexandre Pouget. 2004. “The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation.” *Trends in Neurosciences* 27 (12): 712–19. https://doi.org/10.1016/j.tins.2004.10.007.

Kwisthout, Johan, Todd Wareham, and Iris van Rooij. 2011. “Bayesian Intractability Is Not an Ailment That Approximation Can Cure.” *Cognitive Science* 35 (5): 779–784.

Lettvin, J.Y., H.R. Maturana, W.S. McCulloch, and W.H. Pitts. 1959. “What the Frog’s Eye Tells the Frog’s Brain.” *Proceedings of the IRE* 47 (11): 1940–51.

Ma, Wei Ji, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. 2006. “Bayesian Inference with Probabilistic Population Codes.” *Nature Neuroscience* 9 (11): 1432–38. https://doi.org/10.1038/nn1790.

Maass, Wolfgang. 1997. “Networks of Spiking Neurons: The Third Generation of Neural Network Models.” *Neural Networks* 10 (9): 1659–71.

Maloney, Laurence T., and Pascal Mamassian. 2009. “Bayesian Decision Theory as a Model of Human Visual Perception: Testing Bayesian Transfer.” *Visual Neuroscience* 26 (01): 147. https://doi.org/10.1017/S0952523808080905.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Marr, David, and Ellen Hildreth. 1980. “Theory of Edge Detection.” *Proceedings of the Royal Society of London B: Biological Sciences* 207 (1167): 187–217.

Miłkowski, Marcin. 2016. “Explanatory Completeness and Idealization in Large Brain Simulations: A Mechanistic Perspective.” *Synthese* 193 (5): 1457–78. https://doi.org/10.1007/s11229-015-0731-3.

Montavon, Gregoire, Mikio L. Braun, and Klaus-Robert Müller. 2011. “Kernel Analysis of Deep Networks.” *Journal of Machine Learning Research* 12 (Sep): 2563–2581.

Newsome, William T., Kenneth H. Britten, and J. Anthony Movshon. 1989. “Neural Correlates of a Perceptual Decision.” *Nature* 341 (6237): 52–54.

Oaksford, Mike, and Nick Chater. 2001. “The Probabilistic Approach to Human Reasoning.” *Trends in Cognitive Sciences* 5 (8): 349–357.

Pérez, Toni, Guadalupe C. Garcia, Víctor M. Eguíluz, Raúl Vicente, Gordon Pipa, and Claudio Mirasso. 2011. “Effect of the Topology and Delayed Interactions in Neuronal Networks Synchronization.” Edited by Matjaz Perc. *PLoS One* 6 (5): e19900. https://doi.org/10.1371/journal.pone.0019900.

Piccinini, Gualtiero, and Carl F. Craver. 2011. “Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches.” *Synthese* 183 (3): 283–311. https://doi.org/10.1007/s11229-011-9898-4.

Pouget, Alexandre, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. 2013. “Probabilistic Brains: Knowns and Unknowns.” *Nature Neuroscience* 16 (9): 1170–78. https://doi.org/10.1038/nn.3495.

Sanborn, Adam N., Thomas L. Griffiths, and Daniel J. Navarro. 2010. “Rational Approximations to Rational Models: Alternative Algorithms for Category Learning.” *Psychological Review* 117 (4): 1144.

Schmidhuber, Jürgen. 2014. “Deep Learning in Neural Networks: An Overview.” *arXiv*, 1404.7828.

Shagrir, Oron. 2010. “Marr on Computational-Level Theories.” *Philosophy of Science* 77 (4): 477–500.

Simon, Herbert A. 1998. “Discovering Explanations.” *Minds and Machines* 8 (1): 7–37.

Sporns, Olaf. 2011. *Networks of the Brain*. Cambridge, MA: MIT Press.

———. 2012. *Discovering the Human Connectome*. Cambridge, MA: MIT Press.

Stüttgen, Maik C., Cornelius Schwarz, and Frank Jäkel. 2011. “Mapping Spikes to Sensations.” *Frontiers in Neuroscience* 5. https://doi.org/10.3389/fnins.2011.00125.

Thorpe, Simon J., and Michel Imbert. 1989. “Biological Constraints on Connectionist Modelling.” In *Connectionism in Perspective*, 63–92.

Uhlhaas, Peter J., Gordon Pipa, Bruss Lima, Lucia Melloni, Sergio Neuenschwander, Danko Nikolić, and Wolf Singer. 2009. “Neural Synchrony in Cortical Networks: History, Concept and Current Status.” *Frontiers in Integrative Neuroscience* 3: 17.

Vilares, Iris, James D. Howard, Hugo L. Fernandes, Jay A. Gottfried, and Konrad P. Kording. 2012. “Differential Representations of Prior and Likelihood Uncertainty in the Human Brain.” *Current Biology* 22 (18): 1641–48. https://doi.org/10.1016/j.cub.2012.07.010.

White, J. G., E. Southgate, J. N. Thomson, and S. Brenner. 1986. “The Structure of the Nervous System of the Nematode Caenorhabditis Elegans.” *Philosophical Transactions of the Royal Society London* 314: 1–340.

Yuste, Rafael. 2015. “From the Neuron Doctrine to Neural Networks.” *Nature Reviews Neuroscience* 16 (8): 487–97. https://doi.org/10.1038/nrn3962.

Zednik, Carlos. 2014. “Are Systems Neuroscience Explanations Mechanistic?” In *Preprint Volume for Philosophy Science Association 24th Biennial Meeting*, 954–75. Chicago, IL: Philosophy of Science Association.

———. 2015. “Heuristics, Descriptions, and the Scope of Mechanistic Explanation.” In *Explanation in Biology*, 295–318. Springer.

———. 2017. “Mechanisms in Cognitive Science.” In *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, edited by Stuart Glennan and P. Illari, 389–400. London: Routledge.

———. in press. “Models and Mechanisms in Network Neuroscience.” *Philosophical Psychology.*

Zednik, Carlos, and Frank Jäkel. 2014. “How Does Bayesian Reverse-Engineering Work?” In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, edited by P. Bello, M. Guarini, M. McShane, and B. Scassellati, 666–71. Austin, TX: Cognitive Science Society.

———. 2016. “Bayesian Reverse-Engineering Considered as a Research Strategy for Cognitive Science.” *Synthese* 193 (12): 3951–85. https://doi.org/10.1007/s11229-016-1180-3.