14

Complexity and Computation in the Brain

The Knowns and the Known Unknowns

Karl J. Friston

Abstract

This chapter sets the scene for the treatment of complexity and computation in human cognition and discusses how this treatment is informed by the neurobiological and functional properties of the cerebral cortex. Its agenda is to establish some guiding principles that may help identify hypotheses and computational architectures that go beyond mere descriptions of how the cortex underwrites the repertoire of functions we enjoy, such as action, perception, cognition, affect, and consciousness. In short, it explores the computational imperatives that form the basis for human experience. Complexity and computation are considered, as is how they organize our approach to neuronal dynamics. Criteria are identified that any tenable theoretical framework must respect. In addition, it discusses computational theories that can be entertained, and the degree to which they account for empirical data from anatomy and neurophysiology. Finally, some of the deeper issues that face sentient artifacts are considered that, ultimately, possess a sense of self, purpose, and agency.

Introduction

The purpose of this chapter is to review the fundaments of complexity and computation in the brain and provide some pointers that frame other contributions in this volume. It may seem an almost impossible task to survey all the issues that attend action, perception, cognition, and consciousness in the human brain; however, there are some relatively straightforward principles that make our job much easier. We will pursue the basic theme of complexity and computation, considering carefully what these notions entail. This paves the way for a broad ontology of theories that can be separated into normative theories of *what* the brain is doing and process theories of *how* the brain implements

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

normative imperatives. This separation is useful because it divides the conceptual (*what*) from the empirical (*how*) labor, and allows us to specify clearly the pressing questions that need to be answered.

This chapter comprises four sections. In the first, we will consider the nature of complexity, from the point of view of dynamical systems and self-organization, as well as from the perspective of inference and statistics. This section leaves us with an outstanding issue: How does dynamical complexity relate to structural complexity and vice versa? The second section turns to the notion of computation and the principles that could shed light on computation in the brain. In brief, we will consider computation from the point of view of inference and how this can be grounded to give a physics of computation that can be meaningfully applied to neuronal systems. The third section looks at prevalent normative theories of brain function with a special focus on currently dominant paradigms, such as predictive coding, the Bayesian brain, and active inference. We review these approaches in the light of preceding discussions on complexity and computation. Having addressed the normative side of the challenge, we then consider the more challenging issues of identifying process theories that are consistent with the principles of computation and endorsed by our growing knowledge of cortical and subcortical networks in the brain. This discussion is organized around two scales: large-scale connectomes and hierarchical architectures in the brain, which contextualize smaller-scale processing (e.g., the canonical cortical microcircuit). In the final section, outstanding issues are raised that largely turn on the remarkable capacity for human retrospection and epistemic planning. This, in turn, presents some key questions about the timing of representations and the representations of time. It is at this point that some of the known unknowns start to rear their heads. In other words, this chapter ceases to be a review of what we know and becomes a prospectus for future discussion and work.

Complexity in the Brain

The origin of the word complexity derives from the Latin word *com* (meaning together) and *plex* (meaning woven). A complex system is therefore characterized by its dependencies and interactions, where characteristic, complex behavior *emerges*. This emergence is sometimes taken to mean that there are no high-order instructions or principles that prescribe the interactions—interactions that are generally considered to be "greater than the sum of their parts." However, as we will see later, this is probably not true. Complexity is itself a complex issue, famously reflected in the fact that there is no single definition of complexity. Having said this, in the physical sciences, there are several formal measures of complexity, depending upon the field of application.

Some common examples include computational complexity, usually cast in terms of minimum description lengths that allow people to classify

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

computational problems by complexity class (e.g., P, NP). This is closely related to Kolmogorov complexity and minimum message length in algorithmic information theory (Hinton and Zemel 1993; MacKay 1995; Wallace and Dowe 1999). These are important measures that relate closely to statistical complexity which will play a key role later (Hinton and van Camp 1993). In statistical mechanics and probability theory, complexity is sometimes associated with the notion of entropy; however, this is a slightly naive assumption and misses the point that complexity is really about relationships (i.e., the dependency of entropy over different partitions), such as hierarchical entropy measures over time or the complexity measures that underpin integrated information theory (Tononi et al. 1994). This sort of complexity speaks to the complicated statistical dependencies among the states of the system in question. We will refer to this as *structural complexity* to distinguish it from the *dynamical complexity* that emerges from a system's dynamics.

In dynamical systems there are many forms of complicated (Latin: *com* meaning together and *plicare* meaning to fold) behaviors that rest upon attractor manifolds that are literally *folded* into some mathematical phase or state space. Three common sorts of complexity in dynamical systems theory are reviewed in Table 14.1. In brief, dynamical complexity usually entails an unpredictable space-filling trajectory that, paradoxically, has an attractor of low measure or volume. To unpack these technical terms, what we are saying here is that if one measures a complex dynamical system and plots its states, as in state space over time, the resulting trajectory traces out a path on an attractor or manifold. The peculiar thing about complex systems is that this manifold or attracting set reaches many corners of state space and yet has a very small volume. This is what is meant by "space-filling with low measure." Essentially, this means that complex systems have attracting sets of states which they visit time and time again; however, their paths through state space are convoluted and unpredictable (in the sense of deterministic chaos or other forms of itinerancy). Furthermore, the attracting states ensure that the system will only be found in a very small number of states, compared with the possible states in which it could be found. In many senses, nearly every system we encounter in daily life is an example of a complex system, ranging from the weather through the behavior of our children to nearly every aspect of our exchanges with the world (e.g., our eye movements). A key feature of complex dynamics is their wandering or *itinerant* aspect.

The importance of itinerancy for brain function has been articulated many times (Nara 2003), particularly from the perspective of computation and autonomy (van Leeuwen 2008). Itinerancy provides a link between exploration and foraging in ethology (Ishii et al. 2002) as well as dynamical systems theory approaches to the brain (Freeman 1994). These approaches variously emphasize the importance of chaotic itinerancy (Tsuda 2001) and self-organized criticality (Bak et al. 1988; Kitzbichler et al. 2009; Deco and Jirsa 2012). Itinerant dynamics also arise from metastability (Jirsa et al. 1994) and underlie important

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

phenomena, like winnerless competition (Rabinovich et al. 2008). For a description of these phenomena, see Table 14.1.

Table 14.1 Phenomena that underlie dynamical complexity.

Phenomenon	Description
Chaotic itinerancy	Chaotic itinerancy refers to the behavior of complicated (usually coupled nonlinear) systems that possess weakly attracting sets, *Milnor attractors*, with basins of attraction that are very close to each other. Their proximity destabilizes the Milnor attractors to create *attractor ruins*, which allow the system to leave one attractor and explore another, even in the absence of noise. A Milnor attractor is a chaotic attractor—onto which the system settles from a set of initial conditions—with positive measure (volume). However, another set of initial conditions (also with positive measure) that belongs to the basin of another attractor can be infinitely close; this is called *attractor riddling*. Itinerant orbits typically arise from unstable periodic orbits that reside in (are dense within) the attractor, where the heteroclines of unstable orbits typically connect to another attractor, or they wander out into state space and then back onto the attractor, giving rise to *bubbling*. In other words, unstable manifolds from saddles (i.e., fixed points attracting in one direction and repelling in another) densely embedded in the attractors become stable manifolds and connect different attractors. This is a classic scenario for *intermittency* in which the dynamics are characterized by long laminar (ordered) periods as the system approaches a Milnor attractor and brief turbulent phases, when it gets close to an unstable manifold. If the number of periodic orbits is large, then this can happen indefinitely. The term *ergodic* is used to describe a dynamical system that has the same behavior averaged over time as averaged over its states. The celebrated ergodic theorem (Birkhoff 1931) addresses the behavior of systems that have been evolving for a long time: intuitively, an ergodic system forgets its initial states, such that the probability that a system is found in any state becomes—for almost every state—the proportion of time that this state is occupied.
Heteroclinic cycling	In heteroclinic cycling there are no attractors, not even Milnor ones (or at least there is a large open set in state space with no attractors); there are only saddles connected one to the other by heteroclinic orbits. A saddle is a point (invariant set) that has both attracting (stable) and repelling (unstable) manifolds. A heteroclinic cycle is a topological circle of saddles connected by heteroclinic orbits. If a heteroclinic cycle is asymptotically stable, the system spends longer and longer in a neighborhood of successive saddles, producing a peripatetic wandering through state space. The resulting heteroclinic cycles have been proposed as a metaphor for neuronal dynamics that underlie cognitive processing (Rabinovich et al. 2012) and exhibit important behaviors such as winnerless competition, of the sort seen in central pattern generators in the motor system.

Table 14.1 (continued)

Phenomenon	Description
Multistability and switching	In multistability, there are typically a number of classical attractors which are stronger than Milnor attractors in the sense that their basins of attraction not only have positive measure but are also open sets. These attractors are not connected; they are separated by a basin boundary. However, they are weak in the sense that the basins are shallow (and topologically simple). System noise is then required to drive the system from one attractor to another; this is called *switching*. Noise plays an obligate role in switching but is not a prerequisite for heteroclinic cycling; noise acts to settle the excursion time around the cycle onto some characteristic timescale. Without noise, the system will gradually slow as it gets closer and closer (but never onto) the cycle. In chaotic itinerancy, the role of noise is determined by the geometry of the instabilities. Multistability underlies much of the work on attractor network models of perceptual decisions and categorization; for example, in binocular rivalry (Theodoni et al. 2011).

We have focused on dynamical complexity using concepts that are usually applied to autonomous dynamical systems; that is, systems with dynamics that do not depend upon any independent (i.e., control) variable from outside the system (including time itself). Clearly, to drill down on any particular neuronal system, especially the cortex, we need to acknowledge that it will respond sensitively to outside influences (and may well show time-dependent effects, such as adaptation). In light of this, it might be important to consider nonautonomous dynamical systems, an emerging branch of applied mathematics (Kloeden and Rasmussen 2011), its application in the field of recurrent neuronal networks (Ørstavik and Stark 1998), as well as the analysis of interactive nonautonomous dynamical systems (Schumacher et al. 2012) and causality (Schumacher et al. 2015). The influence of coupled (and therefore nonautonomous) dynamical systems on each other, via the emergence of things like generalized synchrony may have a fundamental role in coordinating neuronal dynamics (Hunt et al. 1997; Schumacher et al. 2012; Friston and Frith 2015), as we will see below.

Measures of Complexity

The title of this section is actually quite loaded. Thus far we have not yet defined complexity; we have just described ways in which it is manifest or can be measured. This is an important distinction because the measurable characteristics of a phenomenon do not, necessarily afford teleological insight. We know that the brain is complex at many levels. Neuronal dynamics are itinerant, show self-organized criticality and metastability (Deco and Jirsa 2012; Cocchi et al. 2017). Furthermore, the dynamic coordination implied by a universe of biological and neuronal rhythms lends the brain a repertoire of complex dynamics and attracting sets that are possibly unparalleled in the universe (Singer and Gray 1995; von der Malsburg et al. 2010). But does this help us understand how the brain works? To harness complexity in a functionalist or teleological sense, it is useful to consider another form of complexity that is closely related to the algorithmic and computational complexity described above.

Statistical Complexity

In statistics and probability theory, complexity has a very particular meaning. It essentially measures the degrees of freedom of a (statistical or mathematical) description of some phenomena or data. These degrees of freedom are technically measured with something like the Kullback-Leibler (KL) divergence between the posterior and prior probability distributions over the causes of data. To unpack this, we must first assume that the world in which we operate is a world of probability distributions and beliefs. Generally, these are distributions over the causes of data or sensory samples; namely, states of the world "out there." Once we describe things in terms of beliefs, we can then evaluate the change in beliefs induced by a measurement or sensory sample. This change is scored by the KL divergence between the posterior belief—after seeing the data or making a measurement—relative to the prior belief—before seeing the data.

In this setting, beliefs are just shorthand for probability distributions of the sort found in Bayesian statistics, or indeed quantum mechanics. This sort of complexity is an attribute of a belief, model, or hypothesis about the causes of outcomes or measures. This may seem a rather colloquial and restrictive sort of complexity; however, it has a much broader scope of application than one might initially guess. This follows from the fact that nearly all interesting physics (and daily life) reduces to some form of inference or measurement. In fact, from the point of view of quantum mechanics right through to general relativity, everything can be reduced to metrology or measurement (Cook 1994). In relation to algorithmic complexity, this means the imperative for efficient communication, decoding, modeling, or hypothesis testing is to *minimize complexity*; in other words, to account for the causes of our sensory interactions with the universe in terms of short messages of minimum complexity (Wallace and Dowe 1999; Schmidhuber 2010). This is nothing more than Ockham's principle.

Putting the Complexities Together

This brief consideration of complexity poses a rather obvious dialectic. If all the principles of algorithmic and information complexity require complexity to be minimized—for example, the principle of maximum efficiency, the principle of minimum redundancy, and so on (Barlow 1961; Optican and Richmond 1987; Linsker 1990)—why is the world so replete with systems that have clear dynamical complexity? We will leave this as a question to be resolved. In fact, any worthy theory of brain function should be able to resolve this dialectic. Before turning to candidate theories, let us now consider the ground rules for computation.

Computation in Humans and Other Animals

Originally, computers were people who made calculations during the Industrial Revolution (Latin: *com* meaning together and *putare* meaning thinking, or reckoning). This is important because computing was, and always has been, a competence of humans, even if the modern perspective on computation focuses on artificial (*in silico*) computation. Nowadays, computation is nearly synonymous with computer science; namely, any type of calculation that follows a well-defined model that can be articulated as an algorithm or scheme. But does this definition really help us?

Let us take a step back and think about what it means to compute or infer. On this view, one can think of computing in terms of deduction, induction, and abduction.[1] In relation to the *model* that underlies a well-defined computation, this translates into inferring the causes or meaning of some measurements or data through deductive, inductive, or abductive algorithms or reasoning. Most computer science treatments of computation would fall under the class of deductive or inductive (Turing style) computations. From the perspective of complexity and computation in the brain, this sort of computation is relatively uninteresting (because it is predicated on propositional logic, as opposed to dynamics and probability theory). We will therefore assume that the sort of computation that characterizes complex self-organizing systems is abductive in nature (e.g., the implicit algorithms we see playing out in meteorology, natural selection, and human perception). So what is abduction?

Loosely speaking, abduction can be thought of as inference to the best explanation. It is characteristically ampliative, in the sense that it often goes beyond the evidence or measurements at hand. In other words, it describes algorithms that appear to bring more to the table than is intrinsic to the computer's inputs. From a mathematical or statistical perspective, the closest algorithm we have to describe this form of computation is Bayesian belief updating, which calls on prior beliefs to contextualize the likelihood of sensory observations. This enables posterior beliefs to be formed that are an optimal assimilation of

[1] Generally defined, *deduction* is "the deriving of a conclusion by reasoning," whereas *induction* refers specifically to "inference of a generalized conclusion from particular instances." *Abduction* is defined as "a syllogism in which the major premise is evident but the minor premise and therefore the conclusion only probable." The crucial distinction is that unlike deduction and induction, abduction is inherently probabilistic.

current evidence (computational inputs) into past experience (prior beliefs). This may seem to be another colloquial formulation of computation; however, it is difficult to think of any complex self-organizing system that cannot be cast in terms of probabilistic or Bayesian updating. Figure 14.1 provides an example of self-assembly and morphogenesis based purely on system dynamics that implicitly perform an elementary form of inference. This inclusive view of computation can be applied to evolution, motor control, and, perhaps, human experience itself (Ao 2009; Harper 2011; Frank 2012).

Figure 14.1 illustrates self-organization through (subpersonal) computation, based on minimizing variational free energy (an information theoretic quantity that measures the surprise or implausibility of some sensed data, given a model of how those data were generated). This simulation shows how simulated (colored) cells can self-organize into a particular form simply by computing and minimizing free energy: the target morphology is shown in the insert on the bottom right. Each time step in this simulation can be thought of as modeling the migration and differentiation of eight cells over several minutes. Upper panels show the time courses of system states encoding cell identity (left), the associated system states mediating migration and signal expression (middle), and the resulting trajectories, projected onto the first (vertical) direction and color-coded to show differentiation. These trajectories progressively minimize free energy (lower left panel), resulting in a differentiation of the ensemble (lower middle panel): the softmax function of the cells' internal states can be interpreted as the posterior beliefs; each cell (column) occupies a particular place in the ensemble (rows); white denotes a probability of one. The lower right panel shows the ensuing configuration: the trajectory is shown in small circles for each time step; the insert corresponds to the target configuration.

To the extent that one subscribes to this formulation of computation, it offers a useful and incontrovertible definition: computation can be defined as *any process that increases model evidence*. Model evidence will be a key concept in what follows and appears in many guises throughout the physical and life sciences. In brief, model evidence is the probability of some data or sensory state of a system, given the system in question. Conceptually, it is useful to treat a system and a model as synonymous. This allows one to think of any system as performing some computation on measurements of—or sensory exchanges with—its world. It is this model that lends computation its defining attribute. Examples of model evidence include the wave function in quantum mechanics, whose squared amplitude corresponds to the probability of a particular state given the quantum system or model in question (Ballentine 1970). In statistical mechanics, the negative log evidence becomes a thermodynamic (Gibbs) energy or potential (Ao 2008; Seifert 2012). In statistics per se, model evidence is also known as marginal likelihood (Beal 2003). In information theory, negative log evidence is known as self-information (Jones 1979); namely, the surprise (or surprisal) induced by an unlikely outcome. This definition is important because it means that the average of (negative log) model evidence

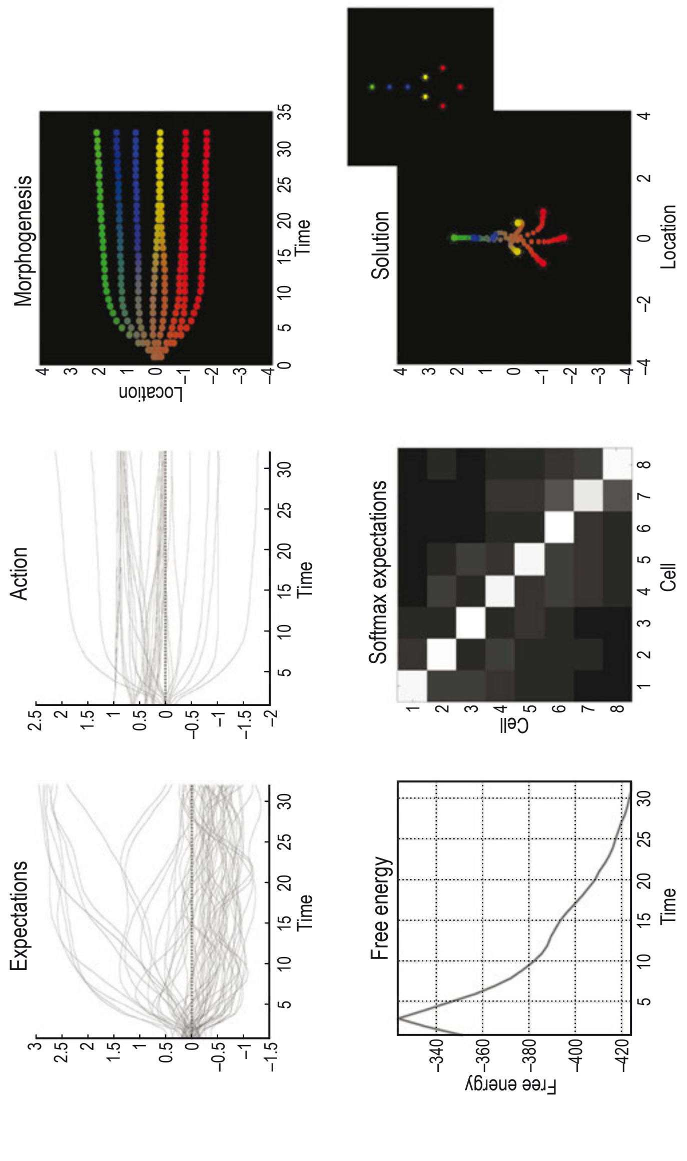

Figure 14.1 Self-assembly and morphogenesis.

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

is entropy. In turn, this means that any self-organizing system that resists the second law of thermodynamics is implicitly minimizing its informational and thermodynamic self-information on a moment-to-moment basis (Nicolis and Prigogine 1977; Friston 2013). Equivalently, interesting self-organizing systems with complex (and complicated) attracting sets must therefore maximize model evidence. So if we define computation as the maximization of model evidence, what does this tell us about the complexity of computation?

Complexity and Computation

This is where the statistical and algorithmic definitions of complexity come into play. Put simply, model evidence can always be expressed as *accuracy minus complexity*. In other words, the evidence associated with some sensory input is just the difference between accuracy (the expected log probability of sensations, given a model of how they were caused) and complexity (the KL divergence between posterior and prior beliefs encoded by the model or system in question). On this view, we can regard self-organization in complex systems as internal responses to external perturbations; namely, sensory inputs. This means that one can associate the internal states of a system with belief states about what is causing or generating its sensory impressions on "the outside." This outside could be a heat bath in statistical thermodynamics (Seifert 2012) or the sensorium in human perception (Still et al. 2012). See Figure 14.2 for an illustration of how complexity minimization underlies action perception in the human brain.

Now let us take another look at the imperative that underlies computation. If computation necessarily increases model evidence, it must therefore entail a decrease in complexity. This is consistent with the minimum message length and algorithmic complexity reduction associated with maximum efficiency and Ockham's principle (Barlow 1974; Hinton and Zemel 1993; Wallace and Dowe 1999). Indeed, some people believe that all self-organization and adaptive behavior can be described in terms of minimizing complexity in one way or another (Schmidhuber 2006, 2010). This brief consideration of computation from the point of view of inference poses two fundamental challenges for any theory of brain function:

- How can we formulate neuronal computations that underlie action, perception, cognition, and consciousness to increase model evidence and the implicitly minimize complexity?
- How does the minimization of *algorithmic complexity* explain the emergence of *dynamical complexity* in sentient, self-organizing systems such as the brain?

To address these challenges, we now consider several global brain theories and see how they fare.

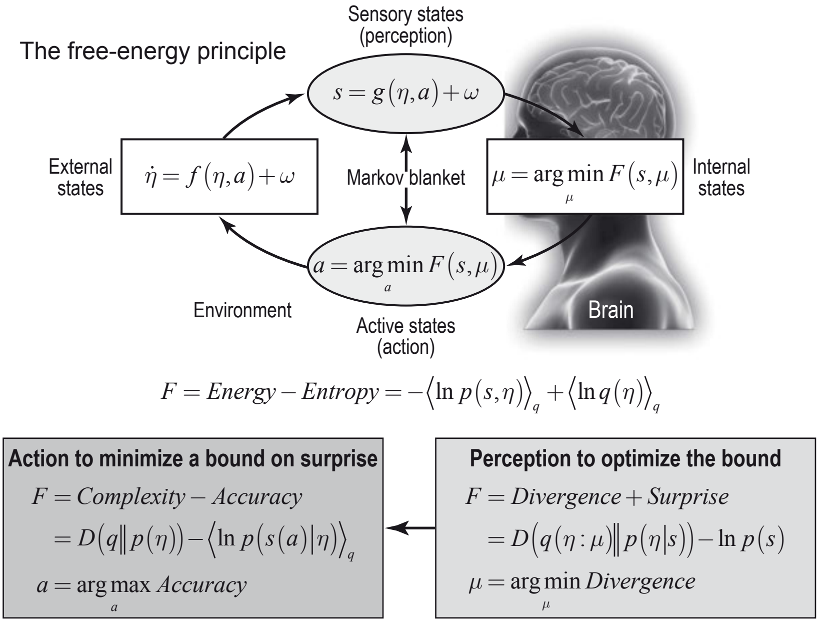
The free-energy principle
Sensory states (perception)
$s = g(\eta, a) + \omega$
External states
$\dot{\eta} = f(\eta, a) + \omega$
Markov blanket
$\mu = \arg\min_{\mu} F(s, \mu)$
Internal states
$a = \arg\min_{a} F(s, \mu)$
Environment
Active states (action)
Brain
$F = Energy - Entropy = -\langle \ln p(s,\eta) \rangle_q + \langle \ln q(\eta) \rangle_q$
Action to minimize a bound on surprise
$F = Complexity - Accuracy$
$= D(q \| p(\eta)) - \langle \ln p(s(a)|\eta) \rangle_q$
$a = \arg\max_{a} Accuracy$
Perception to optimize the bound
$F = Divergence + Surprise$
$= D(q(\eta : \mu) \| p(\eta|s)) - \ln p(s)$
$\mu = \arg\min_{\mu} Divergence$

Figure 14.2 Bayesian computation in the brain. Upper panel: Schematic of the quantities that define a system and its coupling to the world. These quantities include the internal states of a system μ (e.g., a brain) and quantities describing exchange with the world; namely, sensory input $s = g(\eta, a) + \omega$ and action a that changes the way the environment is sampled. The environment is described by equations of motion, $\dot{\eta} = f(\eta, a) + \omega$, which specify the dynamics of (hidden) states of the world, η. Here, ω denotes random fluctuations. Internal states and action both change to minimize free energy or self-information, which is a function of sensory input and a probabilistic belief $q(\eta : \mu)$ encoded by the internal states. Lower panel: Alternative expressions for free energy illustrating what its minimization entails. For action, free energy (i.e., self-information) can only be suppressed by increasing the accuracy of sensory data (i.e., selectively sampling data that are predicted). Conversely, optimizing internal states make the representation an approximate conditional density on the causes of sensory input (by minimizing KL divergence). This optimization makes the free energy bound to self- information tighter and enables action to avoid surprising sensations.

Normative and Process Theories of Computation in the Brain

This section fleshes out the important distinction between *normative* and *process* theories that could be entertained in the neurosciences, with a special focus on the imperatives for complexity and computation described above. How might one approach theoretical frameworks for computation in the brain? Perhaps the easiest thing to do is to distinguish between theories or principles that describe *what* the brain does from process theories that describe *how* the brain does something. We will refer to these as normative (or state) and process theories, respectively.

The perspective in this section borrows heavily from the physical sciences, where it is almost self-evident that to understand any system—from the systems of quantum mechanics through to the canonical ensembles of statistical thermodynamics—it is necessary to specify the system's *Lyapunov function*. The notion of a single (Lyapunov) function that can describe the entire behavior of any system—indeed the universe—may seem fantastical; however, nearly all physics ultimately falls back on some form of Lyapunov function. Put simply, one can describe any (random) dynamical system—whether it is complex, self-organizing, or not—in terms of a set of (random) differential equations (Tomé 2006; Ao 2008; Seifert 2012). Furthermore, the flow or changes in the state of the system at any point in state space (i.e., for any given state) can be completely described by a (Lyapunov) function of those states.

Common examples here include the Lagrangian of gauge theories (e.g., general relativity) (Capozziello and De Laurentis 2011; Sengupta et al. 2016), the Hamiltonian of classical mechanics, the thermodynamic free energies of statistical mechanics, and the Schrödinger Hamiltonian of quantum mechanics (Ballentine 1970; Seifert 2012). All of these quantities are essentially the same thing and just score the improbability of occupying a particular state, such that the flow of the system will tend to vacate regimes of state space in which it is not typically found (i.e., which do not constitute parts of its attracting set). The physical analogies here are not terribly important. The important point is that a complete description of any system can be obtained if we understand what the system is doing in terms of the Lyapunov function it is continuously trying to minimize. The insight here is that we can describe any complex self-organizing system (including the brain) in terms of an apparent *optimization*. This is because—in virtue of decreasing its Lyapunov function—the brain will appear to be optimizing the Lyapunov function. So what is the Lyapunov function for the brain? The answer is exactly the same answer for any system: the self-information or negative log evidence.

There is a long and technical (and very interesting) back story to this assertion; however, we will simply accept this to be the case and recognize some common instances of the implicit self-evidencing implied by this formulation. First, self-evidencing, as the label suggests, implies that systems like the brain are in the game of maximizing the Bayesian model evidence for their models of the sensorium (Hohwy 2016). This is nothing more or less than the Bayesian brain hypothesis cast in terms of a state or normative theory (Ballard et al. 1983; Knill and Pouget 2004; Yuille and Kersten 2006). This has a long history dating back to the students of Plato, through Kant and Helmholtz (Helmholtz 1866/1962) to modern-day formulations in terms of perception as hypothesis testing and variational formulations, such as the free-energy principle and active inference (Gregory 1980; Hinton and Zemel 1993; Dayan et al. 1995; Friston 2010). Implicit optimization also subsumes dominant theories in psychology (e.g., reinforcement learning) and in economics (e.g., expected utility theory). In both reinforcement learning and expected utility theory, the underlying premise is that there is some reward, cost, value, or expected utility function that behavior is trying to realize. In active inference, this function is the expected model evidence over (prior) preferences about outcomes in the future (Friston et al. 2015b; Mirza et al. 2016). I will return to this later; for the moment, let us first see whether this account meets the challenges posed above.

Algorithmic and Statistical Complexity: The Dialectic Resolved

If self-evidencing minimizes the complexity of our (generative) models of the world (Hohwy 2016), how does this explain the dynamical complexity of neuronal activity? It turns out that the answer is relatively straightforward. The explanation here has two parts: the first *dynamical* and the second *structural*. In terms of dynamical complexity, minimizing self-information or maximizing log evidence necessarily engenders *self-organized criticality* and the three mechanisms that underpin dynamical complexity (see Table 14.1). The reason is subtle but intuitive. If any system is trying to minimize its self-information, we can think of the system as tracing out a trajectory on a (self-information) function over its state space. For example, imagine a steel ball rolling over a curved surface, always searching for the lowest points. It can be seen immediately that the states which the system visits will repeatedly correspond to the (local) minima of the (self-information) surface, thereby creating an attracting set. Clearly, because the number of points that constitute a minimum is far less than the total number of points in state space, this attracting set will have low measure or volume. Now, here comes the clever bit. Because the curvature of this surface determines the precision or confidence of posterior beliefs, its curvature determines complexity. In other words, if the ball represents some neuronal population firing rate that is trapped in a very narrow ravine of the (self-information) surface, the system can be very confident about where it is located. Statistically, this is reflected in things like Fisher information and information geometry (Amari 1998). The important thing here is that as the curvature of self-information increases, the difference between the posterior and prior increases, and the complexity increases. This means that, by definition, regions of state space with low self-information must be relatively flat (much like river estuaries are broader than the hanging valleys from which their tributaries emerge). The flat aspect of these attracting minima means that the ball can roll around the local minima in any unconstrained, slowly oscillating or meandering fashion. These critical slowing and long-range fluctuations are the hallmark of self-organized criticality (Bak et al. 1988; Shin and Kim 2006). Furthermore, because the sites of the self-information minima are relatively shallow, this affords the opportunity to jump from local minima to local minima, thereby affording a mathematical image of metastability and other forms of critical dynamics (see Table 14.1 and Figure 14.3). We will see later that precision is itself a quantity that is optimized by the brain, and this optimization may be what we call attention. This lends attention an interesting

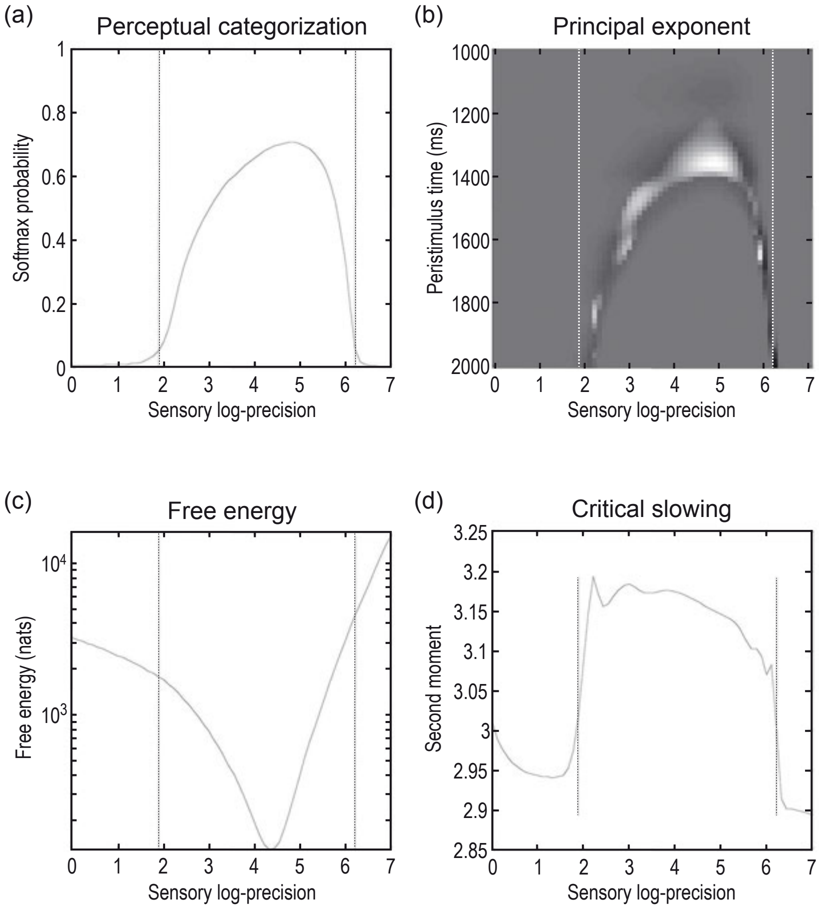

Figure 14.3 Self-organized criticality and computation: (a) The average probability, following stimulus onset, of correctly identifying a song over 64 values of precision on the motion of hidden attractor states. The two vertical lines correspond to the onset and offset of nontrivial categorization—a softmax probability of greater than 0.05. The variation in these average probabilities is due to the latency of the perceptual switch to the correct song. This can be seen in (b), which shows the principal conditional Lyapunov exponent (CLE) in image format as a function of peristimulus time (columns) and precision (rows). It can be seen that the principal CLE shows fluctuations in, and only in, the regime of veridical categorization. Crucially, these fluctuations appear earlier when the categorization probabilities were higher, indicating the prevalence of short latency perceptual switches. Time-averaged free energy is shown in (c) as a function of precision. As one might anticipate, this exhibits a clear minimum around the level of precision that produces the best perceptual categorization. In (d), a very clear critical slowing is shown in, and only in, the regime of correct categorization. In short, these results are consistent with the conjecture that free-energy minimization can induce instability and thereby provide a more responsive representation of hidden states in the world. Adapted from Friston et al. (2012), to which the reader is referred for further details.

interpretation; namely, it might be the psychological homologue of self-organized criticality that allows us to engage selectively with the sensory world in both space and time (Coull and Nobre 1998; Feldman and Friston 2010).

In short, the very mathematical structure of computation in a Bayesian or abductive sense necessarily entails self-organized criticality and fluctuations of the sort that characterizes dynamical complexity. But what about structural complexity and the form of neuronal architectures (e.g., connectivity)?

Generative Models and Structural Complexity

In the search for accurate and minimally complex models of the sensorium, the best solution is generally to recapitulate the causal or statistical structure of the world "out there," within the system (e.g., the brain). In short, the best path to self-evidence is to have a veridical and parsimonious model of the world in which you are navigating. This is an old insight first articulated formally in synergetics in terms of the good regulator theorem (Conant and Ashby 1970; Seth and Friston 2015); namely, any system that can regulate its environment must possess a model of that environment. This tells us something very interesting. It means that minimizing complexity—while maintaining an accurate explanation or prediction of sensory inputs—will cause statistical regularities and causal structure in the world to be transcribed into the system's internal architecture. If one pursues this argument, then we have a natural explanation for the finely crafted and interwoven connectivity in our brains that has all the hallmarks of complexity (Friston and Buzsáki 2016). This is simply a restoration of the sparse, deep, or hierarchical structure of the world "out there," generating sensory impressions. In short, if we live in a complex and complicated world, the minimization of complexity—in the algorithmic sense—not only enforces self-organized criticality and dynamical complexity, it also mandates a structural complexity that mirrors the world. For some, these may be pleasing accounts of complexity and computation in the brain. So is this the end of the story? Clearly not, it is only the beginning. Having a state or normative theory is certainly very useful; however, it says nothing about the process theories that actually explain neuronal computations.

Process Theories

Clearly, there are many theories about neuronal processing that appeal to a greater or lesser extent to normative theories. Happily, the dominant theory—predictive coding—subsumes most available process theories. Predictive coding is not a normative theory; it is a particular algorithm or process theory that has attracted a lot of attention over the past decades in explaining many aspects of neuronal anatomy, physiology, psychophysics, and, more recently, motor control (Srinivasan et al. 1982; Rao and Ballard 1999; Friston 2011b). In brief, predictive coding was originally formulated for compressing large files, based on the minimum description length notions above (Elias 1955). In the neurosciences, it now represents the most developed and established process theory for hierarchical message passing in the brain (Mumford 1992; Bastos et al. 2012). Predictive coding is just an algorithm or scheme that minimizes self-information or maximizes model evidence by updating internal states (i.e., representations) in the light of sensory evidence. It is distinguished from other formulations by calling on auxiliary variables termed *prediction errors*. Prediction errors are simply the difference between sensory inputs (or intermediate representations at hierarchical levels in hierarchical predictive coding) and predictions of those inputs based on internal representations or expectations. In the brain, predictive coding is usually described as reciprocal message passing among the levels of the cortical and subcortical hierarchy (Friston 2010). The recurrent aspect of this message passing is important and fundamentally asymmetric. In other words, top-down or descending messages convey predictions of expectations in the level below (or sensory input per se), whereas ascending or bottom-up signals communicate newsworthy prediction errors that update expectations, thereby improving predictions and resolving prediction errors throughout the hierarchy.

In engineering, the Kalman filter (a special linear case of Bayesian filtering) is the formal homologue of predictive coding. Predictive coding in its generalized form also provides a nice metaphor for several other important schemes used for data assimilation and uncertainty quantification; for example, reservoir computing and deep learning (Schmidhuber 2006; Hinton 2007; Tenenbaum et al. 2011; Salakhutdinov et al. 2013; LeCun et al. 2015). To understand this, we have to distinguish between *inference* and *learning*. In this chapter, inference corresponds to the estimation of (time-varying) causes in the world that are generating sensations, whereas learning corresponds to accumulating experience in the service of updating (time-invariant) model parameters that underwrite inference. Happily, when we put prediction errors into the algorithmic mix, things like *backpropagation of error* can be implemented using Hebbian or associative plasticity. Furthermore, schemes like reservoir computing and liquid state machines (Maass et al. 2002; Buonomano and Maass 2009) can also be considered as variants of predictive coding; for a discussion of how reservoir computing can self-organize to improve predictive coding, see Toutounji and Pipa (2014). The twist here is that instead of optimizing the parameters that enable predictive coding schemes to make better predictions, parameters mapping from a reservoir of dynamics are optimized to select the best prediction of temporally fluctuating inputs, or some supervised output. This sort of scheme (based on recurrent neural networks) has found a particularly powerful application in neurorobotics, reproducing many lifelike behaviors (Tani 2003; Tani et al. 2004).

Generalized predictive coding schemes also provide a nice vehicle for many other issues that attend the dynamic coordination of message passing in the brain (von der Malsburg et al. 2010). A key example here is the encoding

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

of uncertainty through the precision or gain afforded by prediction errors. Technically, this corresponds to the Kalman gain or precision in Bayesian filtering formulations of evidence accumulation or assimilation (Clark 2013). Physiologically, there are a host of important mechanisms that may coordinate the implicit gain control of prediction errors. These range from the control of classical modulatory neurotransmitter systems through to excitation–inhibition balance in the coupling between superficial pyramidal cells and inhibitory interneurons (Yu and Dayan 2005). This is a particularly fascinating area that has clear correlates (of selective evidence accumulation) in terms of fast synchronized neuronal oscillations, which may be a crucial aspect of gating and communication in perceptual synthesis (Singer and Gray 1995; Fries 2005; Womelsdorf et al. 2007; Giraud and Poeppel 2012).

Oscillatory dynamics may also be a key player in process theories of forward and backward message passing in hierarchical predictive coding. The implicit mathematical structure of this message passing suggests that faster fluctuations in prediction errors may be communicated by high frequencies, whereas lower frequencies may convey descending connections (Bastos et al. 2012, 2015). If true, this puts the nonlinear integration of units encoding expectations and prediction errors within the same cortical column center stage in cortical computations (Kopell et al. 2011; Lee et al. 2013a).

The Functional Anatomy of Predictive Coding

So how does predictive coding fare as a process theory in relation to anatomy and physiology? Its explanatory scope is impressive. For example, it provides a principled explanation for the functional asymmetries between ascending and descending (forward and backward) extrinsic (between-area) connections in cortical hierarchies (Mesulam 1998; Hilgetag et al. 2000). These functional asymmetries entail the spectral asymmetries in neuronal oscillations above and established dissociations between driving (forward) and modulatory (backward) synaptic effects (Sherman and Guillery 1998, 2011; Bastos et al. 2012). Variants on different proposals for the integration of hierarchical or centrifugal patterns of extrinsic connections have emerged over the past few decades, starting with the seminal work of David Munford (1992) on the computational architecture of the neocortex. This work has been refined and embellished over the years, leading to detailed descriptions of canonical microcircuits for predictive coding that identify computational roles for individual cell types (Bastos et al. 2012; Shipp 2016); see also Figure 14.4, where the equations in the left panel provide a mathematical form for predictive coding and emphasize the key role of precision (see above) in coordinating and contextualizing the impact of prediction errors on belief updating. As noted above, the way that precision enters into belief updating in these schemes suggests a close link between optimizing precision and attention. In other words, some of the complexity associated with neuronal dynamics rests upon self-organized coupling,

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

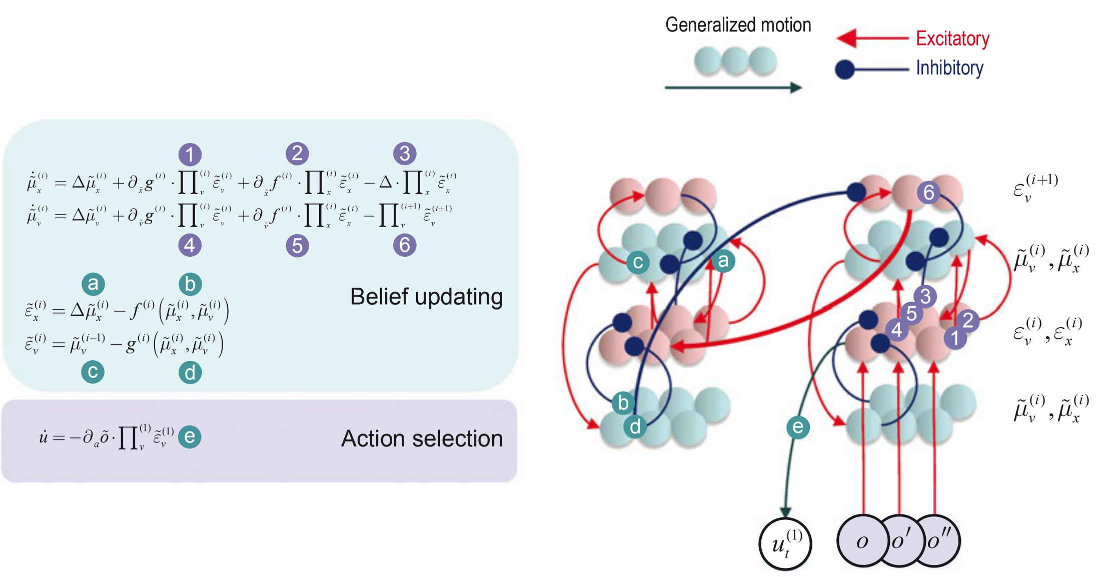
Generalized motion
Excitatory
Inhibitory
1
2
3
$\dot{\tilde{\mu}}_x^{(i)} = \Delta\tilde{\mu}_x^{(i)} + \partial_{\tilde{x}} g^{(i)} \cdot \prod_v^{(i)} \tilde{\varepsilon}_v^{(i)} + \partial_{\tilde{x}} f^{(i)} \cdot \prod_x^{(i)} \tilde{\varepsilon}_x^{(i)} - \Delta \cdot \prod_x^{(i)} \tilde{\varepsilon}_x^{(i)}$
$\dot{\tilde{\mu}}_v^{(i)} = \Delta\tilde{\mu}_v^{(i)} + \partial_{\tilde{v}} g^{(i)} \cdot \prod_v^{(i)} \tilde{\varepsilon}_v^{(i)} + \partial_{\tilde{v}} f^{(i)} \cdot \prod_x^{(i)} \tilde{\varepsilon}_x^{(i)} - \prod_v^{(i+1)} \tilde{\varepsilon}_v^{(i+1)}$
4
5
6
a
b
$\tilde{\varepsilon}_x^{(i)} = \Delta\tilde{\mu}_x^{(i)} - f^{(i)}\left(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}\right)$
$\tilde{\varepsilon}_v^{(i)} = \tilde{\mu}_v^{(i-1)} - g^{(i)}\left(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}\right)$
c
d
Belief updating
$\dot{u} = -\partial_a \tilde{o} \cdot \prod_v^{(1)} \tilde{\varepsilon}_v^{(1)}$
e
Action selection
$\varepsilon_v^{(i+1)}$
$\tilde{\mu}_v^{(i)}, \tilde{\mu}_x^{(i)}$
$\varepsilon_v^{(i)}, \varepsilon_x^{(i)}$
$\tilde{\mu}_v^{(i)}, \tilde{\mu}_x^{(i)}$
$u_t^{(1)}$
o
o'
o''

Figure 14.4 Canonical microcircuits for predictive coding. This schematic illustrates how the mathematics of predictive coding schemes can be used to develop a detailed process theory for neuronal computation. This example is based on the anatomy of intrinsic and extrinsic connections described by Bastos et al. (2012). Left panels show the update dynamics for a hierarchical generalization of predictive coding, where the outputs of a higher level are the hidden causes of state dynamics in the level below. In this hierarchical setting, the prediction errors include prediction errors on both hidden causes and states. In the right panel, prediction errors have been assigned to granular layers that receive sensory afferents and ascending prediction errors from lower levels in the hierarchy, along with superficial pyramidal cells that broadcast ascending prediction errors. Prediction errors are denoted by ε, expectations by μ, and precision by ∏. The functions ***f*** and ***g*** embody a deep (i.e., hierarchical) generative model of sensory observations *o* that are solicited by action *u*; Δ is a differential operator. Adapted from Friston et al. (2017a), to which the reader is referred for further details.

which may entail synchronous gain (Singer and Gray 1995; Fries et al. 2001b, 2008; Fries 2005; Bauer et al. 2006, 2014; Womelsdorf et al. 2007; Bendixen et al. 2012; Auksztulewicz and Friston 2015; Wildegger et al. 2017).

So, if predictive coding appears to capture so much of the brain's computational anatomy, do we have a complete picture (at least at a mesoscopic scale) of computation in the human brain? I would submit that we probably do not. In the final section, let us thus turn to some of the deeper challenges that constitute the focus of much current research.

Beyond Predictive Coding

There are many reasons to suppose that predictive coding in and of itself is an incomplete process theory for neuronal computations. The most obvious shortcoming is its failure to account properly for action, planning, or intentions (Bernier et al. 2017). This problem can be finessed, in part, by appeal to active inference which, essentially, equips predictive coding schemes with classical reflexes (Friston et al. 2015b). This renders motor control a problem of predicting the proprioceptive consequences of action and speaks to purposeful behavior in terms of planning and inference (Attias 2003; Botvinick and Toussaint 2012; Mirza et al. 2016). This is an important extension of predictive processing and hierarchical Bayesian inference in the brain; however, it may raise more issues than it resolves. We will briefly consider a few of these key issues which are considered in other chapters in this volume. We start with some basic aspects of generative models that underlie purpose and selfhood and then consider some of their implications for neuronal dynamics.

Temporal Thickness and Counterfactual Depth

Current trends in machine learning may be taken as a pointer for developments in computational neuroscience, exemplified by the success of deep convolution networks and deep learning (LeCun et al. 2015). This direction, however, is probably not fit for purpose for several reasons. First, deep *learning* and associated reinforcement learning paradigms do not address *inference*. In other words, although data is accumulated in the service of optimizing connection strengths, the problem of how hidden states of the world are inferred online (i.e., data assimilation) is, most often, not in the remit of machine learning. This speaks to the fact that there are usually no dynamics involved in the classification problems addressed in machine learning. In other words, the problem of recognizing static images of handwritten digits is very different to the problem of anticipating the intention and motor behavior of somebody writing digits by hand.

Second, current machine learning approaches using deep neural networks do not usually consider the encoding of uncertainty and, more importantly, epistemic value and intrinsic motivation (Ryan and Deci 1985; Oudeyer and Kaplan 2007; Schmidhuber 2010; Friston et al. 2015b). In other words, if the brain is in the game of minimizing (expected) self-information, it is technically trying to minimize uncertainty. This follows because the time average of self-information is entropy (Jones 1979). The minimization of uncertainty through explorative behavior or epistemic foraging is probably one of the most important imperatives for neuronal computation; thus, it has to be an integral part of any normative theory. Unfortunately, most machine learning algorithms do not accommodate this, leading to ad hoc ways of resolving the exploitation–exploration dilemma (Cohen et al. 2007). In contrast, to understand fully how the brain recognizes and acts upon cues with epistemic affordance, we need to have a clear understanding of how the brain models the future, even in simple tasks like reading (Figure 14.5). Reading is particularly interesting because it speaks to the assimilation of sensory evidence at multiple hierarchical or deep temporal scales (Poeppel et al. 2008), while at the same time calling upon efficient foraging of the visual scene for salient information (Hassabis and Maguire 2007; Mirza et al. 2016). This also brings us to the notion of mnemonics and counterfactual depth which underlie purposeful and possibly mindful behavior (Palmer et al. 2015; Seth 2015).

Deep Temporal Models

If we choose our behaviors based upon a model of the world, then that model must entertain the consequences of action in the future. Furthermore, this future must be encoded at different temporal scales. This is literally a deep and intriguing problem that has clear implications for the form of generative models embodied by neuronal connections and neurophysiology. Again, language and reading provide excellent opportunities to understand how narratives are synthesized in the brain and contextualize our active sampling of the sensorium (Barlow 1974; Beim Graben et al. 2008; Poeppel et al. 2008; Giraud and Poeppel 2012; Dehaene et al. 2015; Konig and Buffalo 2016). There are some fascinating issues when it comes to the details of the underlying process theories (O'Keefe and Recce 1993; Buzsáki et al. 2013; Murray et al. 2014; Friston and Buzsáki 2016):

- How do we select models?
- How do we integrate the assimilation of sensory evidence over different timescales?
- Do we have a moving temporal frame of reference?
- Are there separate spatial and temporal representations (Dehaene et al. 2015), or do we represent dynamic spatiotemporal trajectories?
- How do we contextualize our evidence accumulation and action selection?
- Do we have separate perceptual and motor representations or are these fundamentally integrated within generative models of the sensed world (Grafton and Hamilton 2007; Bernier et al. 2017; Cogan et al. 2017)?

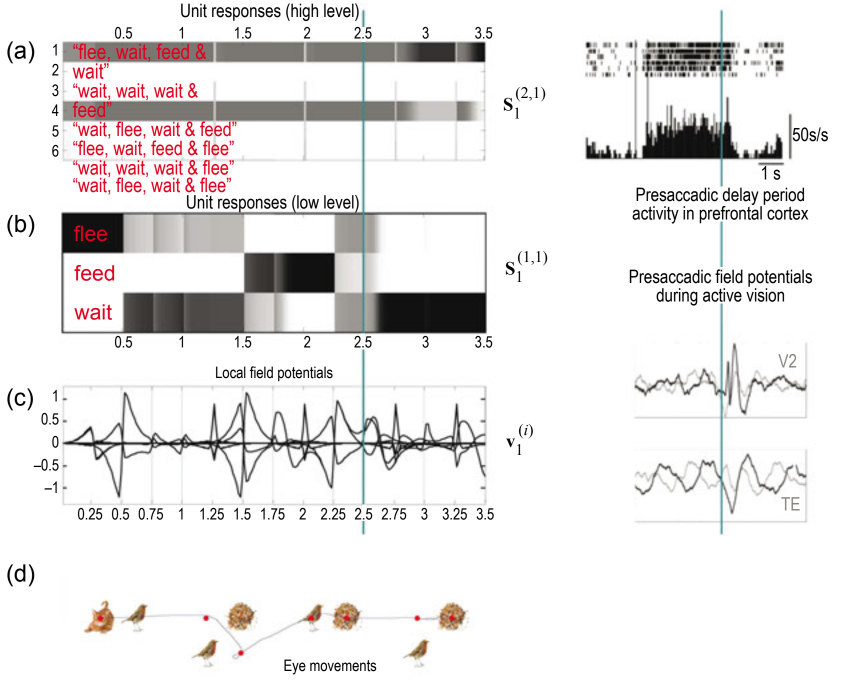

Figure 14.5 Simulated electrophysiological responses during reading: this figure illustrates the sort of Bayesian belief updating that underlies reading. Neuronal dynamics are shown in terms of expectations about hidden states of a deep model (i.e., a model with hierarchical depth) generating words from sentences and pictograms from words. The upper panels show expectations are shown at the higher (a) and lower (b) hierarchical levels in raster format, where an expectation of one corresponds to black (i.e., the firing rate activity corresponds to image intensity). The horizontal axis is time over a reading trial, where each iteration corresponds roughly to 16 ms. The vertical axis corresponds to six sentences at the higher level and three words at the lower level. The resulting patterns of firing rate over time show a marked resemblance to delay period activity in the prefrontal cortex prior to saccades. Saccade onsets are shown by the vertical (cyan) lines. The inset on the upper right is based upon the empirical results reported in Funahashi (2014). The transients in (c) are the simulated firing rates in the upper panels filtered between 4 Hz and 32 Hz, and can be regarded as (band pass filtered) fluctuations in depolarization. These simulated local field potentials are again remarkably similar to empirical responses. The examples shown in the inset are based on the study of perisaccadic electrophysiological responses during activation reported in Purpura et al. (2003). The upper traces come from early visual cortex (V2), while the lower traces come from inferotemporal cortex (TE). Eye movement trajectories produced in this simulation of active inference are shown in (d). Adapted from Friston et al. (2017b), to which the reader is referred for further details.

- How does coevolution make the job of predicting the sensed world easier? Indeed, how do we take turns in predicting each other (see Ghazanfar and Takahashi 2014)?

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 what extent does our ability to predict rely upon neuromodulatory mechanisms, such as dopamine and noradrenaline, or indeed a level of consciousness (Strauss et al. 2015)?
- What is the role of brain oscillations (Mayer et al. 2016), such as theta-gamma coupling, between the prefrontal cortex and hippocampus?

These questions are particularly prescient in the context of predictive coding and language processing (see Arnal and Giraud 2012; Giraud and Poeppel 2012).

A clear utility of deep temporal models is the ability to understand neuronal dynamics in terms of message passing and belief updating. In other words, this formulation provides an explicit account of computations that are "about something." This top-down approach, starting from the generative model and then unpacking the requisite dynamics, can be contrasted with bottom-up approaches, such as deep learning, and the use of recurrent or unfolded networks that aspire to the same kind of functionality. However, these "black box" approaches do not necessarily admit the same level of algorithmic or functional interpretability. For a taxonomy of computational architectures that speaks to the recurrent neural networks used in machine learning and predictive coding-like schemes, based upon generative models, see Harris et al. (this volume).

Finally, it should be noted that the metaphor offered by predictive coding is only appropriate when dealing with continuous state spaces in continuous time. While this is perfectly fine for luminance contrast and perhaps visual motion as well as the detailed kinematics of muscle movements, it may be the wrong sort of parameterization for the lived world. In other words, our intentions, concepts, and sense of self usually come along as (sequences of) discrete or categorical states (see Dehaene et al. 2015; Wilson et al. 2017). Technically, these questions call for a completely different set of neuronal computations that inherit some similarities from predictive coding but are quintessentially different in their form. This is not a problem; indeed, it speaks to the known unknowns that can guide empirical research. For example, there are very particular predictions based on the belief propagation under discrete models (e.g., Markov decision processes) in comparison to continuous models (e.g., the state space models of predictive coding). There is clearly no right or wrong answer in terms of process theories, which means that empirical data will, ultimately, be in a position to adjudicate among the different hypotheses. To do this, however, one must be able to specify the dynamics and implicit coordination implied by various process theories. This is the challenge.

Conclusion

In summary, we have taken the basic nature of complexity and computation to consider the constraints on state or normative theories of brain computation, with a special focus on self-evidencing and inference as the most promising formulation. From this, a number of different process theories—illustrated here with predictive coding—inform, or are informed by, empirical study and highlight the known unknowns in neuroscience. Further questions remain to be addressed regarding subconscious and conscious inference, minimal selfhood, interoceptive inference, emotions, and related areas in philosophy. Many ideas are emerging under the notion of the self-organizing and self-evidencing brain that promise to enrich this enquiry.

Acknowledgments

K. J. F. is funded by the Wellcome Trust (Ref: 088130/Z/09/Z).