Computational Complexity and the Function-Structure-Environment Loop of the Brain

Brendan Juba

Washington University in St. Louis

1 Brookings Dr.

St. Louis, MO 63130

bjuba@wustl.edu

Abstract

At present, the brain is viewed primarily as a biological computer. But, crucially, the plasticity of the brain's structure leads it to vary in functionally significant ways across individuals. Understanding the brain necessitates an understanding of the range of such variation. For example, the number of neurons in the brain and its finer structures impose inherent limitations on the functionality it can realize. The relationship between such quantitative limits on the resources available and the computations that are feasible with such resources is the subject of study in computational complexity theory. Computational complexity is a potentially useful conceptual framework because it enables the meaningful study of the family of possible structures as a whole—the study of *"the brain,"* as opposed to some particular brain. The language of computational complexity also provides a means of formally capturing capabilities of the brain, which may otherwise be philosophically thorny.

Keywords: computational complexity, circuit complexity, neural networks, invariance, identifiability, learning theory

1 Introduction

The brain is generally understood to be a kind of computer. As a first, gross simplification, it is generally believed that the brain takes sensory inputs and produces motor commands as outputs, for the purpose of prolonging the survival of its host's genome. Although this view can help orient the study of the brain, we would naturally like to understand its function more precisely. We possess a pretty thorough picture of the behavior of the key cells comprising the brain, in particular neurons (Koch, 1998; Koch and Segev, 2000; London and Häusser, 2005; Herz et al., 2006); but, while clarity at the cellular level is necessary to understand the human brain, it is insufficient to understand it. This is because, as we will review in Section 2.1, unlike most other organs, the function of the brain seems particularly sensitive to the arrangement and behavior of even a handful of cells. The cellular characterization of the brain's behavior is therefore inadequate to provide an understanding of its function for two reasons, that we will develop in Section 3. The first, widely acknowledged reason, is that the human brain at the cellular level is therefore too large and complex a system for us to comprehend in full detail. The second reason is less acknowledged, but is in a sense more fundamental: the human brain is therefore crucially sensitive to the idiosyncracies of individual development and physiology, which we review in Sections 2.2-2.3. It may therefore be incorrect to speak of "the" human brain as a single system, in the same way one might discuss "the" human eye or "the" human respiratory system. Instead, our object of study is the range of functional behaviors of the family of brains that develop in a closed-loop interaction with natural environments. We can seek to address both difficulties through the appropriate use of *abstraction.*

1.1 The theory of algorithms informs the study of the brain

The first difficulty stems from our inability to understand the behavior of more than a few cells at a time. Addressing this difficulty thus involves some abstraction of the behavior of such groups of cells as performing some simpler computational function. The use of such abstraction in studying

the human brain was advocated in a classic work by Marr (1982). Marr advocated the characterization of algorithms and representations as a bridge between the behavior of individual neurons and a precise characterization of the brain's behavior at a functional level. A description of the brain's function in terms of algorithms and representations gives us a narrative in which we can place the actions of the individual neurons as providing such representations, or as executing such algorithms.

The study of the human brain at this level of algorithms and representations is the focus of the field of cognitive science. But, such algorithms and representations are studied more generally by computer science. Indeed, the analogous separation of function from implementation in silicon is recognized as crucial in enabling humans to engineer the highly complex computer hardware and software we use today. In computer science, the study of algorithms and representations is divorced from the requirement in cognitive science that these algorithms capture the particular, natural systems of the brain. This freedom has enabled computer science to develop a reasonably successful suite of techniques for the analysis of such systems, in order to facilitate their design. I will argue that some of these techniques should also be useful in understanding the brain at this abstract level.

1.2 The role of computational complexity

In particular, I will argue that the framework of *computational complexity* provides a conceptual orientation and some basic mathematical techniques to help overcome the second difficulty; that the study of "the" human brain is actually the study of the many possible brains that may result from development and experience. We must analyze something more general than a particular brain, or, perhaps, even more general than a particular algorithm. One such kind of analysis, that I will describe in Section 4, draws on the observation at the cellular level that human brains, and structures within these brains, tend to have roughly similar sizes, densities, signaling speeds, and so on. These measurable, quantitative constraints on a typical brain constrain the algorithms that the

brain can implement, especially on short time scales. A constraint on the algorithms in turn constrains the possible functions that the brain could be computing. Such relationships are the objects of study in computational complexity—the relationships between the *quantitative "complexity" constraints* of an implementation level in some computational model and the *possible computational behaviors* realized at the abstract, functional level. Specifically, computational complexity addresses questions about whether or not a function can be computed at all within some given quantitative constraints, and about the relative power, that is, relative breadth of functions computable under different constraints and models.

Actually, a surprising outcome of these studies has been that many different computational models, when allowed a reasonable amount of resources (most notably computation time), have essentially the *same* amount of power, that is, compute exactly the same functions. In particular, we do not know of any currently feasible model of computation more powerful than the Turing machine[1], for example, but we know of many models of equivalent power. This state of affairs motivates the *Strong Church-Turing Thesis:* roughly, that the notion of feasible computation is largely insensitive to the choice of model. Or, in other words, the Turing machine model in particular, when subject to a reasonable constraint on its running time, expresses precisely the family of functions that can be feasibly computed. This hypothesis enables a mathematically precise study of efficient computation *in general.*

In turn, such a study of efficient computation in general enables a kind of "quantitative philosophy" that provides a second application for computational complexity to the study of the brain. In short, the language of computational complexity allows us to sensibly formulate questions about the high-level abilities of the brain. For example, in Section 5 I will sketch an example explication of "predictive ability." This is enabled by a precise complexity-theoretic formulation what

[1] A Turing machine is a simple, idealized computer, given by a finite control unit with an unbounded length memory tape. The details will not be significant to us, but it behaves as follows. On a single step of the machine, it reads and updates the contents of a single symbol on the tape, updates the state of the control unit, and moves to an adjacent cell of the tape. The "time" taken by a Turing machine corresponds to the number of such steps.

constitutes a "reasonable" encoding of semantic information content. Note that such questions are philosophically thorny otherwise: in the absence of a known, fixed formal representation scheme, who is to say what was computed and what was not? In particular, there is no single, fixed function (at the highest of Marr's three levels) that captures such abilities in general. Nevertheless, the notions of computational complexity provide a way of capturing such things.

The value of such a definition is that by abstracting away the details of the encoding used by a particular system, it is then possible to ask precisely whether or not such capabilities are possessed by a given region of the brain, the brain of another species, or a computer program. Perhaps most crucially of all, when a system does *not* have such abilities, the mathematical techniques of computational complexity may provide a means to demonstrate their absence.

2 Functional and structural variability in the brain

It is hard to understate the functional significance of the structural variation across brains. To our current understanding, all of the differences in skills, knowledge, culture, memories, personality, and so on between individuals are the result of differences in the fine structure of their brains. Even so, it is somewhat surprising how much of the brain's structure, even at larger scales, seems to be determined by the environment in which it develops, as opposed to having been genetically determined. The consequences of this extreme sensitivity and variability motivate turning to a complexity-theoretic model of the brain, so we will review the evidence for this view of the brain.

2.1 Learning and the sensitivity of neural representations

It is widely understood that the brain has the capacity to learn. This learning, in turn, must be realized by some kind of mechanism in the brain. One mechanism for learning was first proposed by Hebb (1949). Recall that much of the brain consists of a network of cells called *neurons,* joined at *synapses*. Neurons are capable of producing an electric *spike*, that can be detected across these

synapses by other neurons, sometimes causing them to generate spikes as well if the overall level of activity is high enough. In Hebb's mechanism(s), the relative sensitivity of a neuron to spiking activity across its various synapses is modified when the neurons spike simultaneously, or in sequence. Cellular mechanisms capable of supporting Hebb's proposal were subsequently observed, starting with work by by Bliss and Lømo (1973). Althought it was plausible that learning in the brain was the result of Hebb's mechanism for learning by altering these synaptic weights, Yang et al. (2009) recently confirmed the existence of another mechanism for learning: they observed the creation of new dendritic spines in mice, leading to the creation of *new synapses*, as well as the elimination of existing spines, eliminating synaptic connections. In short, both the network of connections among neurons which happen to have their axons and dendrites near one another, as well as the strength of these connections, appear to be subject to continual change. In simple models of neural networks, as for example proposed by McCulloch and Pitts (1943), then, no part of the network can be assumed to remain fixed.[2]

This *plasticity* of the brain is perhaps surprising when one considers that it appears that the firing of just a handful of neurons may represent specific content. Quian Quiroga et al. (2005) describe their observation of a *"Halle Berry"* neuron during recordings of the spiking activity of neurons in a neurological patient: the particular neuron in question spiked when the patient was shown a picture of the actress Halle Berry, or when the patient was shown the words "Halle Berry," but not apparently in response to any other stimulus the patient was shown. It thus appears that the firing of one or a few neurons really does encode the concept of "Halle Berry" in this particular patient's brain. Quian Quiroga et al. were able to find a variety of other cells that responded selectively to various images, and Quian Quiroga et al. (2007) argue that this suggests that these concepts are represented by the firing of a population of a small number of neurons. Others (Bowers, 2009) argue more strongly that these findings support the "grandmother cell" hypothesis, essentially that

[2]Valiant (2015) suggests a more liberal view of a *potential synapse* as occurring wherever axons and dendrites pass near enough to one another that a synapse could potentially be created. It is still plausible that the network given by these potential synapses does remain fixed.

information is represented by the firing of *single* neurons in the brain.

Either way, it does appear that the kind of local changes in the brain's structure that have been observed can radically alter the functional behavior of these networks. Hence, the brain seems particularly sensitive to variation at the cellular level among individuals, and even within the same individual over time during development and learning. It therefore seems necessary to study the function and capabilities of the brain at a more abstract level than the actual cellular network.

2.2 The role of environmental factors in shaping architectural features of the brain

A classical view of the brain holds that the various skills and functions of the brain are localized in small regions. Beyond the evident features of the brainstem, cerebellum, hippocampus, etc. that can be identified with various functions, a vast array of sensory and motor tasks involved in vision, language, planning, etc. are believed to be carried out by various small patches of the cortex. Much of this view of *localization* was first put forth by Brodmann (1909). These classical views were largely based on the experience of loss of function of various kinds by patients who suffered damage to one part of their brain or another. Most famously, the first identified case was Broca's patient Tan Tan. Tan Tan, apparently due to a lesion in the front of the left hemisphere of his cortex, could only speak one syllable, "tan." His gestures led Broca and others to conclude that he could understand the language of others well enough, though. Thus, the natural hypothesis was that the damaged area was linked to language production. Modern *in vivo* imaging studies of healthy, functioning brains lend support to this view: for example, Moro et al. (2001) observed that activity in "Broca's area," an anatomical feature approximately located where Tan Tan suffered his injury, can be used to *detect* grammatical errors. More strikingly, a subsequent study by Musso et al. (2003) suggests that activity in this area may be tied to the grammar of natural language in particular, not simply rule learning in general. Concretely, subjects were taught constructed languages containing rules that violate the model of Universal Grammar that captures all known human languages; activity in Broca's area then predicted correct performance for rules that are

consistent with Universal Grammar, and predicted performance errors for the inconsistent rules.

In view of the observed phenomenon of localization, a natural first hypothesis would be that the various functions of the cortex develop in portions of the cortical tissue in a genetically determined way. Such a course of development according to a genetically-determined schedule would be much like how the other organs of the body differentiate from stem cells during development. Indeed, the observed pattern of connectivity of the cortex, which aided Brodmann's division of the cortex into regions, lends some weak support to such a view.

Strikingly, genetics *per se* appears *not* to be the primary factor determining the function of the cortex. Sur et al. (1988) determined that the auditory regions of ferret pups could be induced to develop features typical of the visual regions by surgically moving the retinal projections from the region of cortex usually associated with vision to the auditory regions. Sur et al. were thus led to posit that the it is the inputs to a region of cortex during development that determines its function. That is, the visual areas develop in response to being connected to optic nerves, the motor regions develop in response to the thalamus and spinal cord, and so on. This view is further supported by the observed recruitment of the unused cortical regions by the remaining sensory modalities in congenitally blind (Cohen et al., 1997) and deaf (Bavelier and Neville, 2002) humans.

This is roughly in line with Lashley's doctrine of *equipotentiality* (Lashley, 1929): that any portion of the cortex has the capacity to carry out the function of any other. Equipotentiality, like localizationism, was originally supported by observations of patients with damaged or missing portions of their brain; but in this case, damage early in development, which leads to no apparent lasting deficiency (in contrast to damage to the developed brain).[3] A beneficial upshot of such hypotheses is that studies of the algorithm controlling development in one region of the cortex should transfer to others. A problematic upshot is that the results of such a developmental process

[3] Along similar lines, motivated by observations of the structure and behavior of different regions of cortex, Mountcastle (1978) proposed that the various, hierarchically arranged regions of the cortex all employed the same algorithm for learning. The mechanisms by which cortical regions develop their unique character during development seems fundamentally distinct from the mechanisms underlying other kinds of learning, however (Greenough et al., 1987).

are likely to vary across individuals; and, on account of the apparently localized representation of information in the cortex (cf. Section 2.1), these variations likely have significant impact on the way these fundamental computations are performed across the brains of various individuals.

2.3 Variability in interconnections

So far, we have observed that it appears that any significant structural feature of the cortex *can* change—be it at the level of weights, synapses, or even functional regions. Moreover now, beyond the variability of the network of synapses in the cortex due to this plasticity, we observe that it appears that even the basic scaffolding of interconnections formed during development is not fixed. Braitenberg and Schüz (1998) describe *Peter's rule,* that this network of interconnections is formed during development by the idiosyncratic pattern of contact between cells. The pattern of interconnectivity in this network is therefore essentially random. Indeed, even the density of interconnections along specific pathways has been found to vary widely across individuals (Hilgetag and Grant, 2000). So, in summary, not only are the weights and synapses of the brain formed by the idiosyncratic processes of learning and development, but moreover the *initial* configurations and larger scale aggregate structure of these networks are also highly diverse.

3 Sufficiently rich computational models are underdetermined

Everyday experience tells us that the human brain is capable of learning to address a wide variety of tasks. Since the brain's function is apparently determined by its current, conditioned structure, it is perhaps unsurprising that the structure of the brain varies at all levels across individuals, and that the (apparently) localized representation of information in the brain renders the functions computed by portions of the brain highly sensitive to even small variations in this structure.

This poses a major barrier to the accurate modeling the brain of a single individual, let alone a population of individuals. Although it is surely possible to estimate various statistics of the distri-

bution of such variability, we should expect that this diverse range of variation across individuals will be reflected by a functionally significant variation in the lower-level architecture of the brain, specifically, in the architecture of regions of the cortex. We should not expect the corresponding cortical regions of two individuals to compute the same function. And, to the extent that they compute similar functions, we should not expect them to be computed in the same way. So, we should not expect to be able to determine the functional architecture of the brain (across individuals) on the basis of such statistics alone.

The development of theoretical models capturing the workings of a brain is further hindered by the inherent weakness of the link between functional behavior and the underlying model whenever this model is computationally rich. In the first place, the underlying model itself is not determined by the functional behavior. In cognitive psychology, this is known as *non-identifiability*.

I will review these issues and examine the prospects for addressing them, e.g., by measurable "performance characteristics." Ultimately, in response to these difficulties, I will suggest in Section 4 that we should model "the" brain (or, precisely, regions of the brain) as belonging to a *family* of possible *functions* rather than a precise computational architecture.

3.1 Invariance and non-identifiability

The algorithms and representations used by the brain, i.e., at Marr's intermediate level (Marr, 1982), are studied in cognitive psychology. In a classic work, Anderson (1978) argued that the functional, stimulus-response behavior of subjects is inherently inadequate to determine key aspects of algorithmic architecture. That is, the underlying algorithms and representations used by the brain are not *identifiable* on the basis of the observed competence (or non-competence) of subjects at various tasks. Anderson's observation is closely related to the various forms of the Church-Turing thesis: essentially, his claim follows from the ability of one algorithmic architecture to simulate the functionality of another. The Church-Turing thesis is based on the observation that such simulations yield many models capturing precisely the same family of functions. The

strong version of the Church-Turing thesis is based on the observation that whenever a natural model of time complexity exists, these simulations incur an overhead that is limited by some polynomial function of the running time and input size; so, as long as we allow for some polynomial "slack" in the amount of time, again we find that many natural models of feasible computations are also equivalent. (As we will review later, however, we *can* exhibit differences between the models if we take a more refined view, and don't allow arbitrary polynomial slack.)

A simple example of such non-identifiability occurs in circuits. Suppose we have a circuit of gates computing the Boolean AND, OR, and NOT functions, where the AND and OR have arbitrary numbers of inputs: recall that the AND function outputs 'true' if *all* of its inputs carry 'true,' and outputs 'false' otherwise, whereas the OR function outputs 'false' if *all* of its inputs carry 'false,' and outputs 'true' otherwise. So the OR and NOT gates can *simulate* the AND gate: we can exchange the role of 'true' and 'false' by negating all of the inputs to the OR gate using NOT gates, and then negating its output (this is known as *de Morgan's law*). So, we cannot identify the difference between a network consisting of AND, OR, and NOT gates, and a network consisting only of OR and NOT gates (or AND and NOT gates). Notice, this simulation only uses three layers of gates to simulate a single AND gate, so if the circuit consisted of a constant number of layers ℓ (supposing for simplicity that there are no outputs feeding back to a higher layer), the simulation can be carried out in 3ℓ layers. More radically, consider the NAND function which outputs 'false' if all of its inputs are 'true' and outputs 'true' otherwise. This gate computes the NOT function on a single input, and so can actually be used to simulate both the AND function using two layers of gates, as well as the OR function using two layers of gates; thus, an AND/OR/NOT circuit that is ℓ layers deep can be simulated by a NAND-only circuit that is 2ℓ layers deep and vice-versa. Likewise, the total number of gates in this NAND-only circuit only needs to be twice that of the original AND/OR/NOT circuit, (plus possibly the size of the input) since we only need to either compute a negation of the output of a gate, or compute the negation of the gate feeding its input.

Moreover, now, suppose we consider the family of *linear threshold* gates: on n inputs $x_1, \ldots, x_n$,

these are parameterized by real-valued weights $w_1, \ldots, w_n$, and a threshold θ. Supposing 'true' corresponds to 1 and 'false' corresponds to 0, the linear threshold gate outputs 'true' when $\sum_{i=1}^{n} w_i x_i \geq \theta$ and outputs 'false' otherwise. Such gates are rich enough to express AND, OR, and NAND (and many other functions besides). Therefore, there may be many different threshold circuits that ultimately express precisely the same Boolean computation in different ways. Such threshold gates correspond to the McCulloch-Pitts model of a neuron (McCulloch and Pitts, 1943). Therefore, at least in the McCulloch-Pitts model, we see that we will encounter difficulties if we wish to determine the actual architecture of a circuit based on its input-output behavior alone. Edelman (2004) discusses such systems, in which in particular there are multiple, functionally equivalent structures expressible by the same model. He calls such systems *"degenerate."* Since, as we reviewed in the previous section, the development of the connectivity and weights of the neural circuitry of the cortex appears to be highly idiosyncratic, we would certainly expect to encounter a wide variety of the possible circuits in this particular degenerate system.

The models of computation by our different kinds of gates are particularly similar; it is possible for far more diverse models to simulate one another, with correspondingly more serious consequences. Work on modeling the feedforward path of the ventral stream of the visual cortex, carried out independently by Serre et al. (2005), and Thorpe et al. (Thorpe, 2002; Thorpe et al., 2004), serves as a mild cautionary tale. These two teams of researchers gave quantitative models of the function of the visual cortex constructed from biophysically plausible circuits, and demonstrated via computer simulations that the models so constructed performed well on standard vision tasks. This is a significant achievement, as it is a substantial and necessary test of any would-be theory of the workings of the brain. Consequently, both teams seemed to feel confident that they had proposed a model that would likely become the starting point for developing "the" model of the visual areas, and ultimately the entire brain. Imagine, then, the surprise of the researchers to discover that another team had achieved similar results utilizing quite dissimilar underlying mechanisms! The major difference between the two approaches was how data was encoded: Serre et al. had used a

more traditional spike rate based encoding scheme, whereas Thorpe et al. used a temporal scheme, i.e., one in which earlier spikes have higher weight. Thorpe (2002) initially argued for a temporal coding based on the difficulties that a rate-based scheme would encounter in accounting for the fast response times observed in practice, and Guyonneau et al. (2005) have performed a more extensive theoretical study suggesting that the timing of spikes is what shapes a neuron's response. On the other hand, Serre et al. (2005) have tested the performance of their model against the performance of human subjects extensively, and found that it does fairly well at predicting the performance of the human subjects on visual tasks.

The lesson we should take away from this incident is that Marr's hierarchy (Marr, 1982) is not merely a taxonomy. The tasks of identifying the physiology and identifying the functionality are quite distinct, and should not be confused. The fact that two rather different underlying models were able to produce similar results should not surprise us so much. It follows from our prior discussion of equivalences across models of computation that a given class of problems will often have many equivalent formulations. The difference from Edelman's notion of degeneracy is that he is considering equivalent structures in a fixed family of models (e.g., the family of McCulloch-Pitts circuits), whereas here we are concerned with the possibility that two different families of models – e.g., temporal versus rate-based models here – of the underlying physiology could yield equivalent behavior. So, while the underlying model is critical from the standpoint of neurophysiology and often important from the standpoint of experimental design, studying the functional capabilities will not help us identify which of a set of computationally equivalent models best describes the underlying physiology. So, we must be careful not to claim too much about what such studies say about the physiology. Likewise, if we manage to successfully identify the computational power of the cortex with a class of circuits, further work on identifying the physiology that "truly" yields the computational model (at Marr's lowest level) will generally not contribute to our understanding of its function (at Marr's highest level). The functionality essentially never identifies a unique model, to say nothing of the underlying physiology that (in a sense) implements the model.

3.2 Limitations of model fitting

So, in general, one cannot identify "the" computation underlying a given behavior. But surely, one might think, it should be possible to at least capture the behavior itself, irrespective of whether or not we understand how it is computed. This is essentially the problem of *(supervised) machine learning*: given a collection of example inputs and outputs, can we find the function? Or more precisely, can we at least find a function that is close to the actual behavior, close enough that it can predict behavior under novel circumstances? Of course, we expect the actual functional behavior of individuals' cortical regions to vary, so the function we learned would be unique to that individual, somewhat analogous to sequencing their genome.

The work of Sur et al. (1988) suggests that the brain itself is learning these functions, so perhaps we as observers ought to be able to learn them, too. Indeed, one might be optimistic, given not only the simulations of Thorpe et al. (Thorpe, 2002; Thorpe et al., 2004) and Serre et al. (2005), but also related, recent work involving *"deep neural nets."* These models are more loosely inspired by the observed architecture of the cortex, but have moreover sometimes matched human-level performance in tasks such as labeling images (Krizhevsky et al., 2012), speech recognition (Dahl et al., 2012), and face detection (Taigman et al., 2014). In light of this, one might be optimistic that the functions being trained by these "deep nets" might actually be that learned by the cortex—that because we are getting similar input-output behavior out of computationally feasible model-fitting processes for the same family of models, perhaps they really do capture functionally equivalent representations. Or at least, that they produce models that are representative of "typical" brains.

These hopes are thoroughly dashed by the work of Nguyen et al. (2015). They show that it is easy to construct, for example, images that appear to a human observer to be a slightly "noisy" picture of one type of animal, but that are confidently classified by the deep net as an entirely different type of animal. Note that these counterexamples are consistent with the claim that the trained deep net matching human level performance, as these images are not *natural*: they feature "noise"

that has been carefully chosen with knowledge of the deep net's weights. Learning algorithms, for example in Vapnik's statistical learning model (Vapnik, 1998) and Valiant's computationally-bounded model (Valiant, 1984), only guarantee that a good *approximation* to the target is learned, as measured by the distribution of the data. That is, learning algorithms only promise correct labels on most of the images that are likely in the natural distribution. They do not and cannot promise to identify the actual function. While we can improve the approximation by training on more data, it is infeasible to collect a training set large enough to guarantee good predictions on the entire range of inputs of interest. So, not only do the representations produced fail to enable any deeper analysis of the actual representations used or other features of the cortical computation, they also cannot be used to make predictions beyond the labels on the training distribution. Thus, as they cannot be used to answer our natural questions, they are poor models of the cortical region.

3.3 Avenues for model identification and their challenges

Modeling the function of regions of the cortex in actual subjects is still the most straightforward approach to studying the brain. We will briefly review its prospects. As these approaches face formidable challenges, this review will motivate a more abstract study of the brain in Section 4.

To begin, we note that the function of such circuits must be considered in the context of natural data: The work of Sur et al. (1988) suggests that neural circuits are developed in response to the kind of inputs they receive. As we noted above, the functions produced by such learning systems are only meaningfully fixed in the context of the distribution of their training data. In particular, this suggests that the context of the signaling inputs received from the rest of the cortex will be essential to studying the function of regions that are not immediately connected to sensory inputs; in other words, in vivo studies are therefore essential.

At a sufficiently small granularity, identifiability is not at issue. For example, we can begin to make meaningful sense of the behavior of a single neuron. The studies of Quian Quiroga et al. (2005) have used in vivo recordings of single neurons to great effect here, giving strong evidence

that concepts are represented by the firing of a few neurons. The trouble, of course, is that it is hard to record from more than a relatively small number of neurons at a time, on the order of thousands, whereas cortical regions may contain hundreds of millions of neurons. Such techniques therefore cannot say much about more abstract computations in the brain, which depend on an understanding of the patterns of activity in neighboring regions. Indeed, note that the successes of this technique were at identifying activity with images of faces, names, and buildings, which are relatively concrete sensory inputs.

While the prospects are dim for obtaining useful information from direct recording of neurons in these more abstract regions of cortex, functional magnetic resonance imaging (fMRI) has been used profitably to obtain information at a smaller granularity than the totality of human behavior. Although it cannot resolve the activity of individual neurons, it can at least give a rough picture of the activity in cortical regions in vivo. For example, this technique was used by Moro et al. (2001) and Musso et al. (2003) to investigate the hypotheses that the Broca-Wernicke areas use a Chomskian Universal Grammar to decode and generate linguistic structure, and moreover, that this grammar computation is distinct from the general rule-learning capacity of the brain. Donoso et al. (2014) likewise used fMRI to produce evidence for activity in the pre-frontal cortex capturing the effectiveness of abstract hypotheses being considered by subjects, and predicting changes between such hypotheses. In particular, they proposed an approximate-Bayesian reinforcement learning model that generates similar activity. While such fMRI studies provide useful tests for such well-developed models, due to the coarseness of the signal they provide, however, they are less useful in the development of models. In particular, because fMRI cannot resolve the actual representation of information by the cortex, it cannot actually determine the functions being computed. We can only check the correlation of the strength of activity with rather specific quantities, such as the probability that a subject makes an error of some kind.

A related approach is to use the pattern of such performance errors and computation time to inform the study of the brain at the algorithmic level (Newell, 1990; Pylyshyn, 1984). Although

we know that one computational model can generally simulate another using similar resources, this equivalence breaks down under sufficient precision. For example, Maass (1984) has shown that the basic, single-tape Turing machine cannot determine whether or not a string of letters is a palindrome without taking time that grows quadratically in the length of the string. By contrast, most of our usual architectures (such as the random-access memory model underlying desktop digital computers) can check whether or not a string is a palindrome in time merely proportional to the length of the string. Thus, in principle, the single-tape Turing machine might be ruled out as a computational model for the brain *if* the amount of time taken by subjects to determine whether or not a string was a palindrome appeared to be subquadratically related to the length. The primary issue that arises is that if subjects did appear to be taking quadratic time, then we essentially learn nothing—it is possible for stronger architectures to utilize a sub-optimal algorithm. Moreover, limits will remain on the extent to which we can nail down the actual architecture; although such separations will allow us to rule out some architectures, not all pairs of models can be separated this way. Some simulations are indeed quite efficient such as the circuit simulations we described earlier. The actual power of the different models is then essentially the same, even with respect to the resources available. We then cannot use such a strategy for model identification.

4 Complexity-theoretic modeling of the human brain

So far, we have seen that the actual functions computed within cortical regions are likely to vary widely across subjects, and that the prospects for resolving examples of these functions in detail are dim. I now argue that the brain – or, more precisely, regions of the cortex – are meaningfully studied as *families* or *"classes"* of functions, constrained by the resources available to compute them – e.g., number of neurons, energy, depth, etc. – and their learnability.[4] Computational complexity is precisely the study of the relationship between such resource constraints and the functions that

[4] A similar argument has been made before, see for example Parberry (1994). But, much more has been discovered in the study of both circuits and learning since that work, and the details of my conclusions will differ.

can be computed within such limited resources; computational learning theory (Valiant, 1984) is, in turn, the study of the learnability of functions as informed by computational complexity. Such families of circuits thus provide a model for regions of the cortex capturing the range of *potential* functional variation of a typical brain. A key benefit of taking this relaxed stance, that we wish to identify a *class of functions* as opposed to a computational model of the brain, is that model-invariance now works in our favor: as the classes of functions are relatively insensitive to the details of the computational models, we expect the mathematical theory to tolerate revisions to the picture of the brain at the cellular level.

4.1 Complexity-theoretic models for short time scales

Our starting point for the modeling of the computation performed by the brain is the classical McCulloch-Pitts model of the neuron (McCulloch and Pitts, 1943). In this model, a neuron takes as inputs the presence or absence of spikes along the synapses of its dendrites, encoded by Boolean values, and computes a linearly weighted vote to decide whether or not to itself produce a spike. That is, if the presence or absence of input spikes are encoded by the Boolean variables $x_1, \ldots, x_n$, there are weights $w_1, \ldots, w_n$ indicating the strength of these synapses, as well as whether they are excitatory or inhibitory; the neuron then decides to fire if $\sum_{i=1}^{n} x_i w_i \geq \theta$ for some real-valued threshold θ. In essence, the McCulloch-Pitts neuron thus computes some *linear threshold* gate.

The brain is then modeled as a circuit comprised by such linear threshold gates. More narrowly, we may model the behavior of a region of the cortex on short timescales (e.g., 100-200msec) by an acyclic circuit of highly limited depth, since on such short timescales information appears to flow in only one direction, and the cascade of spikes can only reach neurons a few synapses away (Thorpe et al., 1996). In the coarse-grained framework of computational complexity, we translate such concrete parameters into relative, asymptotic quantities, taking the actual brain to be a typical member of an infinite *family* of such circuits of different sizes. For example, we may naturally choose to cast these parameters as varying relative to the size of the input (number of neurons or

other cells with projections to the region of interest), "n," that provides an index for the family. We may then take the "very small" depth to be bounded by some (arbitrary) constant independent of n; by contrast, parameters of size similar to the input, such as the number of neurons involved in the computation, would be regarded as having an arbitrary polynomial relationship to n.

This is of course not entirely correct—we know gross limits on the sizes of brains and cortical regions, and so such an asymptotic statement is therefore nonsense, strictly speaking. But, the number of layers through with spikes may pass is quite small relative to the hundreds of millions of neurons that may comprise the cortical regions through which such spikes pass. The asymptotic relationship captures the algorithmically significant relationship that one quantity is very small relative to another, without requiring us to specify precisely how much so.

Moreover, modeling the depth of layering of such circuits as bounded by a constant, independent of the input size, is mathematically convenient. In particular, a major benefit of a model formulated in these terms is that it is *invariant* or robust to many variations in its precise formulation; this is the positive side of non-identifiability. For example, more sophisticated models of the neuron that include computation along the dendritic branches (Koch, 1998; Koch and Segev, 2000; London and Häusser, 2005; Herz et al., 2006) will not alter the computational model we use substantially, although they do generally suggest that a single "layer" of neurons is more aptly modeled as two or possibly three layers of threshold circuits.[5] Since our model is given by threshold circuits with small but otherwise arbitrary numbers of layers, such variation in the exact number of layers will be irrelevant to us.[6] We note that the analog properties of the real neurons comprising these threshold circuits is likewise essentially irrelevant from a computational perspective: Maass and Orponen (1998) have shown that the presence of even small amounts of noise renders such circuits

[5]These layers are determined by the connectivity of the neurons – that is, the first "layer" consists of neurons that can be reached directly from the "inputs," the second layer consists of neurons with synapses taking input from neurons in the first layer, and so on. This layering is *not* meant to correspond to the layered structure of the cortical tissue itself.

[6]This stands in contrast to the implication of such developments on the variety of learning strategies available at the neuronal level, where the ability to jointly update two layers is significantly different.

no more powerful than Boolean threshold circuits.

Such circuits are indeed highly parallel; this parallelism allows significant computations to be performed under such severe timing constraints. But, the need for parallelism is *just another constraint on the functions computed*. We note that the variety of functions that can be computed in parallel under such severe constraints is quite limited relative to a desktop computer. In particular, the aforementioned works of Serre et al. and Thorpe et al. demonstrate that cortical regions can be meaningfully simulated on a computer (Serre et al., 2005; Thorpe, 2002; Thorpe et al., 2004).

We now will seek to identify a *sub*-class of such circuits that is *learnable* by feasible algorithms (at least, polynomial-time bounded). Learnability is actually quite a severe constraint. A result by Klivans and Sherstov (2009) shows that even very simple two-layer circuits of linear threshold functions may not be learnable: even if the output layer threshold gate always computes an AND, any strategy that could learn even a nontrivial approximation to the function computed by such circuits could be used to break encryption schemes that are believed to be secure. A corollary of such results is that although we do not know what representations can be learned by deep nets in practice, we certainly *don't* believe that they can even be trained to fit arbitrary two-layer neural networks, even though their architecture may be capable of representing precisely such networks.

More generally, there is a distinction in learning theory between the functions that a learning algorithm can learn, and the kinds of representations that the learning algorithm produces. Both are generally captured by some class of representations, and the expressive power of the class of representations that the algorithm produces necessarily constrains the power of the functions it can learn. So, for example, if we take our constant-layer linear-threshold gate (McCulloch-Pitts neuron) networks as our representation, then we can only learn functions that can be expressed by such circuits. But, the class of functions that are efficiently learnable is often strictly weaker than the representations produced by their efficient learning algorithms, as appears to be the case here.

So, the class of *all* constant-depth, polynomial-size threshold circuits is likely too rich for our purposes. One biologically-motivated further restriction on the class of circuits was proposed by

Uchizawa et al. (2006): they proposed to consider such circuits with the further restriction that on any input, a relatively small number (logarithmically many) of the model neurons "fires," i.e., relatively few of the threshold gates output 'true.' This restriction is motivated by the observations of Margrie et al. (2002) that the spiking activity in the brain is relatively sparse, and the related *energy conservation principle* proposed by Lennie (2003) on theoretical grounds: namely, a firing rate of just a few percent in each local region of the cortex would be metabolically unsustainable.

Now, the learnability of functions represented by a class of circuits appears to be closely related to the tractability of the mathematical analysis of such circuits, for example, to our ability to establish that some functions are *not* computed by such circuits. An informal connection has long been observed between learnability and such circuit complexity lower bounds, and moreover, work by Fortnow and Klivans (2009) has provided some formal connection. Of course, the tractability of the mathematical analysis of such circuits is of fundamental importance to this program of work in its own right. On this count also, the family of all constant-depth, polynomial-size threshold circuits was known to be too rich, again essentially due to the fact that secure cryptography is believed to be feasible within such circuits (Razborov and Rudich, 1997). Fortunately, however, once the class is further restricted to have limited energy complexity as proposed by Uchizawa et al., such analysis becomes feasible: in particular, Uchizawa and Takimoto (2008) were able to show strong lower bounds for computing a simple function (the Boolean inner product) in this model. We still do not know if this class of circuits is efficiently learnable, but given the observed close connection between learnability and such lower bounds, it appears likely to be learnable. This class is therefore a natural candidate for the functions computed by regions of cortex on short time-scales, i.e., along feed-forward paths.

4.2 Modeling for longer time scales

On time-scales longer than a few hundred milliseconds, there is adequate time for the recurrent (feedback) connections to play a significant role in the functionality. Of course, the backwards

connections may already play a role in shaping the response to stimuli on the feed-forward path (a view we'll revisit in Section 5), but there they cannot respond to the stimulus, which substantially increases the computational power of the circuits: Even in a "constant-depth" model, without the energy restriction of Uchizawa et al. (2006), cyclic circuits can simulate general time and/or memory bounded computation (depending on whether the number of iterations is bounded). And, even with a logarithmic energy complexity bound, such circuits can (at least) simulate finite automata with a number of states quadratically larger than the number of gates of the circuit (Šíma, 2014). Finite automata are strongly believed not to be learnable (cf., Kearns and Valiant (1994)), so this fact may not be of much relevance to the computational power of the learned behavior of the individual regions.

On such longer time-scales, however, we may be more interested in the architecture of connections between the features and regions of the brain, in the spirit of Brodmann's map. In particular, a modern update of such a map is the objective of the human connectome project. We might model the individual regions as computing constant-depth, polynomial-size, logarithmic-energy circuits, imposing the constraint of learnability at the level of the individual cortical regions alone. If we now consider the power of the resulting overall architecture, Šíma's result then gives us some idea of its potential extent.

An alternative, related line of work initiated by Valiant (1994, 2000), has aimed to develop a quantitative computational model of the brain that closely models the quantitative aspects of the underlying physiology while approaching the kind of algorithmic model sought in cognitive science. Of course, in light of the challenges posed by identification at Marr's algorithmic level, Valiant is careful to stress that he is *not* asserting that the algorithms he describes in the neuroidal model are actually used by the brain. Rather, he is only seeking to demonstrate that the model, set with realistic parameters, is adequate to capture the range of behavior exhibited by the brain. A current review of the program and its guiding philosophy is given by Valiant (2014). Work by Valiant (2005) and Feldman and Valiant (2009) has demonstrated that this model can support mul-

tiple basic tasks and a plausible amount of storage while using a plausible degree of connectivity, etc. For example, an analysis in Valiant (2006) observed that these ranges of parameters were consistent with the presence of the kind of sparse concept representations conjectured to underlie the observation of the "Halle Berry" neuron by Quian Quiroga et al. (2005).

As the neuroidal model models the brain as performing computations in a way constrained by the quantitative parameters of the observed physiology, it is in particular a computational complexity-theoretic model. One significant distinction between the program underlying the neuroidal model and the program I am suggesting, however, is that I am proposing somewhat looser quantitative constraints than considered thus far by Valiant. The reason is that I would like to understand the relationships between (our models of) the brain and the models considered in the larger complexity-theoretic study of computation. For example, it is natural to ask how the power of a network of neuroids compares to other computational models, such as the threshold circuits of bounded energy complexity discussed earlier. This is not immediate, as although neuroids have a threshold-firing behavior like threshold gates, they also operate according to a nontrivial state-transition model (e.g., resetting after a refractory period), unlike these gates. The neuroidal model with a given range of parameters naturally has greater fidelity to the actual brain than the model I wish to consider; but, the kinds of simulations used to establish such comparisons rely on a little bit of slack or overhead in the computational constraints they obey. It is thus theoretically convenient to regard models as having "the same" or "similar" power in spite of having to lose some resources in converting from one to the other. We can also typically give these arguments in such a way that concrete bounds can still be extracted for the purposes of testing a theory.

A complexity model such as the finite automata model used by Šíma (2014) that counts the number of states, as opposed to simply bounding the complexity of the algorithm, is likely crucial here. Primarily this is because it allows us to meaningfully discuss computation on finite inputs, by "charging" the finite automaton for the complexity of the algorithm it uses, which is kept separate from the space and time used in the Turing machine model. That is, such a state-size versus gate-

size bound makes concrete, testable predictions about the number of neurons, level of activity, etc. required in order to support a given computation on a given input size. It also respects the fact that the brain's computation only needs to be performed on a limited size input. Although we are currently only aware of this feature of the model giving it additional power under somewhat contrived circumstances, such circumstances do exist.

5 Characterizing functionality using complexity-theoretic concepts

A second role for the language and models of computational complexity in the study of the brain is that it allows us to formulate claims and questions about the kinds of computations performed by systems like brains, while abstracting away the particular details of *how* this computation is performed. In this way, it becomes meaningful to pose such questions across different systems that were *not* apparently designed with any particular semantics in mind, e.g. in comparative biology. I will give an example of such "quantitative philosophy" by sketching a definition of "predictive ability," as proposed to capture "intelligence" (Hawkins and Blakeslee, 2004; Clark, 2013).

5.1 The necessity of abstract and generalized theories

In the study of the function of the brain, following Blum et al. (2005), I claim that it is desirable to define properties such as consciousness or intelligence in an abstract or generalized manner. Such a theoretical description of these concepts would of course be complementary to the actual study of the brain's physiology: in the physical study, certain processes should be observed which could be verified to have the necessary properties to be called "conscious" or "intelligent" in the abstract theory. These theories are desirable because they are *necessary* to decide the presence or absence of such high-level features in other systems. For example, a natural question that often arises is whether members of other species – dogs, octopi, etc. – are conscious or intelligent (and *how* intelligent). A more loaded family of questions concern when a human fetus or infant, or a brain-

damaged patient should be considered conscious, which may have serious legal consequences.

Specifically, if we wished to decide whether or not a member of another species, or an entire ant colony, or a robot is conscious or intelligent, it would obviously be unsatisfying to conclude that since none of these have precisely the kind of primate brains in which the various studies of neural representation have been carried out, none could be conscious or intelligent. Turing's test (Turing, 1950), which was proposed as a means to identify "intelligence" without having to specify precisely what "intelligence" is, also has some related, well-known shortcomings. In Turing's test, a judge converses with a human and a machine (or some other subject), and attempts to tell which is which; if the judge cannot tell the difference with probability adequately greater than chance, then the test subject is declared intelligent. The troubles with the test include its failing of subjects who refuse to participate or, for whatever reason, cannot communicate in the judge's language. Clearly, we need to develop theories that describe what it is about the function of the brain that gives rise to these properties without being tied to the particular realization in the systems we observe. When a subject is intelligent or conscious, these theories should permit a simple demonstration of this property by describing how the required systems are implemented. The theory should also make it possible in principle to demonstrate conclusively that a system is *not* intelligent or conscious, although demonstrating the absence of these properties may not necessarily be so simple. Establishing such a claim might require demonstrating that the system in question is incapable of computing some necessary functions using some lower-bound argument, or some kind of clever experiment. Nevertheless, possession of such abstract definitions allows us to meaningfully consider such questions, which is the first necessary step to resolving them.

5.2 Outline of a complexity-theoretic study of high-level functionality

In our study we wish to isolate the tractable portions of the problem – in this case, the functional or mechanical aspects of our high-level attributes – from the philosophically contentious questions of what an entity really "thinks" or "feels." Koch (2004) gives a nice, analogous discussion of the

distinction between the neural correlates of consciousness that he wishes to study and a philosopher's notion of consciousness. We begin by viewing the brain as a formal system, as circuits from a particular class. We continue by asking what properties of those circuits characterize our concepts like intelligence or consciousness: we are seeking the necessary and sufficient conditions that permit our formal system to exhibit the behavioral or mechanical aspects of these high-level properties. We then may define the high-level properties in terms of these conditions.

For example, I will sketch a first attempt at a complexity-theoretic definition of "predictive ability"; Hawkins and Blakeslee (2004) proposed that such an ability to make predictions about one's environment captures our informal notion of "intelligence." (See Blum et al. (2005) for a similar initial attempt at defining "consciousness.") Their proposal was motivated by Rao and Ballard (1999) and Lee and Mumford (2003), who suggested that the feedback connections in the cortex serve to carry predictions about future inputs. (A similar proposal was fleshed out in more detail recently by Clark (2013).) I will attempt to show how this idea might be translated into a formal definition, illustrating where we expect computational complexity theory to play a role. I stress that this definition is agnostic with respect to the underlying mechanisms and encodings used; in particular, it is *not* necessarily tied to the hierarchical Bayesian frameworks in which the ideas were first developed. For example, the "predictive join" operation in the neuroidal model, proposed by Papadimitriou and Vempala (2015) should satisfy our framework, when applied to the inputs to the model. As we will see, this definition abstracts a level above even Marr's functional level, identifying a property that such *functions* may or may not possess.

We will model time as proceeding in discrete steps. Sequences of one input string for each time step, $x_1, x_2, \ldots, x_t, \ldots$, will then represent an environment; we will later have cause to consider *classes* of environments, which are merely sets of possible input string sequences. We will consider our candidate predicting machine M to have some internal state s_t, which in the language of Turing machines would be the contents of its work tapes or in the language of modern computers would be a dump of the contents of its memory; in a more natural setting, we could think of this as a

"snapshot" of neural activity around one moment in time. Regardless of our terminology, if M is in state s_t and sees input x_t, it may perform some action and updates its internal state to s_{t+1}. We assume that the action and s_{t+1} are determined uniquely by s_t and x_t.

In this setting, we'll take "the ability to make predictions" to mean that at time t, M frequently has access to x_{t+1}. We remark that although one could consider richer notions of "predictive ability," this simple notion is still nontrivial, and is sufficient to illustrate how we proceed in providing such definitions. Now, in order for M to be making predictions about x_{t+1}, it must be updating its internal state so that eventually s_t, together with x_t will contain sufficient information about x_{t+1} that M can do better than merely guess its contents, but we will *avoid* requiring it to store this information in any particular format. Instead, we will say that M is intelligent with respect to some class of environments $\mathcal{C}$ if there is some efficiently (i.e., polynomial-time) computable *deciphering function* D_M such that for any environment $\{x_t\} \in \mathcal{C}$, eventually, $D_M(x_t, s_t) = x_{t+1}$ with frequency "non-negligibly" better than chance—formally, better than chance by an amount that is at least $1/\mathrm{poly}(|x_t|, |s_t|)$, that is, some inverse polynomial in the encoding length (in bits) of the inputs and states. We may regard such deciphering functions as generalizing the "read-out" classifiers of Hung et al. (2005). By allowing a computation that extracts (or "translates") the relevant information packed into s_t, the use of a custom deciphering function grants our theory the desired robustness to the details of the machine model, information encoding, and so on.[7] A machine M for which such a deciphering function exists has sufficient information about x_{t+1} in such a format that we may consider it to possess such a prediction, whereas any realistic machine for which no such deciphering function can be implemented could not be doing much better than blindly guessing about the next input, so satisfying this property is necessary and sufficient for the machine to make predictions about the next input from environments in $\mathcal{C}$.

Something like "non-negligibility" is generally necessary for us to even observe a statistical difference in the behavior from random guessing. The requirement of a "non-negligible" advantage

[7]The strong Church-Turing thesis asserts that the definition of "efficiently computable" is similarly robust.

also helps guard the definition against declaring M able to predict on the basis of a deciphering functions with a few special rules that obtain a small advantage by looking at x_t directly. On account of the limitation that D_M be fixed, this generally only offers a negligible advantage (in the technical sense) against rich families of environments. Moreover, we note that in the related case of learning, a result of Freund and Schapire (Schapire, 1990; Freund and Schapire, 1997) establishes that several such "weak predictions" can be used to make "strong predictions" in which the predictions are correct with probability that are arbitrarily close to certainty. In other words, the ability to make such weak predictions (on examples weighted by the performance of the other weak predictors) addresses the algorithmic essence of the learning task.

We remark that one feature of this definition is that it is possible to classify the *degree* of intelligence of a machine M by the richness of environment classes that it can successfully predict; it is easy to see that on one end of the spectrum, any machine is able to predict a constant environment, but no machine is able to predict (the class of) all environments. We are particularly interested in the behavior of the machine with respect to "natural" environments, by which we mean informally the class of environments corresponding to the entity's natural environment, encoded appropriately as Boolean strings. We may speculate that such environments, if captured formally, would be produced by a process featuring an infinite "unobserved state" and each symbol in its state at a given time-step would be computed from portions of its state in the previous time-step that were at most some bounded distance away—a locality constraint in its update rules, but more work would need to be done before we would be satisfied with our definition. Part of the Hawkins-Blakeslee notion of "intelligence" would involve characterizing these environments more precisely; machines capable of predicting such "natural" environments would be considered "intelligent."

We should require that any model of the low-level mechanisms in the brain (e.g., formal models of neural circuits) be sufficiently powerful to exhibit intelligence, defined in this way. Note that we assume that M's storage is bounded, and hence that M cannot simply offload the task to D_M by storing everything. So, the problem of preparing to make such predictions is nontrivial in general,

and this definition does require the models that satisfy it to have some computational power.

Returning to our broader goals, we would like to separate which characteristics of the brain's function should be considered necessary, such as making predictions in the proposal above, and which merely serve to implement those functions. Notice that this definition makes no mention of any details of how the brain carries out the task of making predictions. Rather, it abstractly characterizes or specifies what makes a machine "capable of making predictions." This is analogous to separating the specifications of the neural correlates of consciousness (Koch, 2004), etc., from their implementations. Once this has been accomplished, then we can say that an entity having implementations of the specified functions has the high-level attribute described—intelligence, consciousness, etc.

We expect such definitions to be relevant since this functional description of our high-level properties are precisely what we need for the sorts of abstract and generalized theories discussed earlier. By separating the specifications of the functions from their implementations, we permit implementations in different settings to be constructed or discovered, in particular in other species. In addition, if such theories have been developed and the special case of brains have not yet been fully understood, then we would hope that our definitions would help explain how the function of the brain could give rise to at least the observable aspects of consciousness, intelligence, etc., by clarifying precisely what sort of functionality we are looking for and establishing what kind of underlying models would be necessary or sufficient for exhibiting such functionality.

Acknowledgements

I thank Sashank Varma and Frank Jäkel for a series of conversations that shaped the section on identifiability and invariance, and their comments on a previous draft. I thank Erick Chastain for much helpful feedback on an early version of this work that appeared in the Carnegie Mellon University Undergraduate Research Journal, *Thought* (Juba, 2006). I also thank Leslie Valiant

for an informative discussion on neural modeling, and Manuel Blum, Ryan Williams, and Matt Humphrey for the many influential discussions while working on the CONSCS project. Parts of this work were supported by the National Science Foundation as part of the Aladdin Center (www.aladdin.cmu.edu) under grant CCR-0122581, and associated REU funding. The preparation of the present manuscript was supported by an AFOSR Young Investigator award.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, 85(4):249–277.

Bavelier, D. and Neville, H. J. (2002). Cross-modal plasticity: Where and how? *Nature Reviews Neuroscience*, 3(6):443–452.

Bliss, T. V. and Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate areaof the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology (London)*, 232:331–356.

Blum, M., Williams, R., Juba, B., and Humphrey, M. (2005). Toward a high-level definition of consciousness. In *20th IEEE Computational Complexity Conference*, San Jose, CA. Invited Talk. http://www.cs.cmu.edu/~mblum/research/pdf/CONSCS8.ppt.

Bowers, J. S. (2009). On the psychological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, 116:220–251.

Braitenberg, V. and Schüz, A. (1998). *Cortex: Statistics and Geometry of Neuronal Connectivity*. Springer, Berlin, second edition.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Großhirnrinde*. Barth-Verlag, Lepzig.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36:181–204.

Cohen, L. G., Celnik, P., Pascual-Leone, A., Corwell, B., Faiz, L., Dambrosia, J., Honda, M., Sasato, N., Gerloff, C., Catala, M. D., and Hallett, M. (1997). Functional relavance of cross-modal plasticity in blind humans. *Nature*, 389:180–183.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42.

Donoso, M., Collins, A. G. E., and Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, 344:1481–1486.

Edelman, G. M. (2004). *Wider than the Sky: the Phenomenal Gift of Consciousness*. Yale University Press, New Haven, CT.

Feldman, V. and Valiant, L. G. (2009). Experience-induced neural circuits that achieve high capacity. *Neural Computation*, 21:2715–2754.

Fortnow, L. and Klivans, A. R. (2009). Efficient learning algorithms yield circuit lower bounds. *Journal of Computer and System Sciences*, 75:27–36.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Greenough, W. T., Black, J. E., and Wallace, C. S. (1987). Experience and brain development. *Child Development*, pages 539–559.

Guyonneau, R., VanRullen, R., and Thorpe, S. J. (2005). Neurons tune to the earliest spikes through STDP. *Neural Computation*, 17:859–879.

Hawkins, J. and Blakeslee, S. (2004). *On Intelligence*. Times Books, New York, NY.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, NY.

Herz, A. V. M., Gollisch, T., Machens, C. K., and Jaeger, D. (2006). Modeling single-neuron dynamics and computations: A balance of detail and abstraction. *Science*, 314(5796):80–85.

Hilgetag, C.-C. and Grant, S. (2000). Uniformity, specificity and variability of corticocortical connectivity. *Philosophical Transactions of The Royal Society B*, 355(1393):7–20.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866.

Juba, B. (2006). On the role of computational complexity theory in the study of brain function. *Thought (Carnegie Mellon University Undergraduate Research Journal)*, 1:32–45.

Kearns, M. and Valiant, L. (1994). Cryptographic limitations on learning Boolean formulae and finite

automata. *J. ACM*, 41:67–95.

Klivans, A. R. and Sherstov, A. A. (2009). Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12.

Koch, C. (1998). *Biophysics of computation – Information processing in single neurons*. Oxford University Press, New York, NY.

Koch, C. (2004). *The Quest for Consciousness: a Neurobiological Approach*. Roberts & Company Publishers, Englewood, CO.

Koch, C. and Segev, I. (2000). The role of single neurons in information processing. *Nature Neuroscience*, 3:1171–1177.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pages 1097–1105.

Lashley, K. (1929). *Brain mechanisms and intelligence: A quantitative study of injuries to the brain*. Chicago University Press, Chicago, IL.

Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America*, 20(7):1434–1448.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, 13:493–497.

London, M. and Häusser, M. (2005). Dendritic computation. *Annual Review of Neuroscience*, 28:503–532.

Maass, W. (1984). Quadratic lower bounds for deterministic and nondeterministic one-tape Turing machines. In *Proc. 16th Annual ACM Symposium on Theory of Computing (STOC)*, pages 401–408.

Maass, W. and Orponen, P. (1998). On the effect of analog noise in discrete-time analog computations. *Neural Computation*, 10:1071–1095.

Margrie, T. W., Brecht, M., and Sakmann, B. (2002). In-vivo, low-resistance, whole-cell recordings from neurons in anaesthtized and awake mammalian brain. *Pflugers Archive*, 444(4):491–498.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, New York.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of ideas immanent in neural activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

Moro, A., Tettamanti, M., Perani, D., Donati, C., Cappa, S. F., and Fazio, F. (2001). Syntax and the brain: Disentangling grammar by selective anomalies. *NeuroImage*, 13:110–118.

Mountcastle, V. B. (1978). An organizing principle for cerebral function: The unit module and the distributed system. In *The Mindful Brain*, pages 7–50. MIT Press, Cambridge, MA.

Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., and Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, 6:774–781.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*.

Papadimitriou, C. and Vempala, S. (2015). Cortical learning via prediction. In *Proc. Conference on Learning Theory (COLT 2015)*.

Parberry, I. (1994). *Circuit Complexity and Neural Networks*. MIT Press, Cambridge, MA.

Pylyshyn, Z. W. (1984). *Computation and Cognition*. MIT Press, Cambridge, MA.

Quian Quiroga, R., Kreiman, G., Koch, C., and Fried, I. (2007). Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends in Cognitive Sciences*, 12(3):87–91.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435:1102–1107.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87.

Razborov, A. A. and Rudich, S. (1997). Natural proofs. *Journal of Computer and System Sciences*, 55:24–35.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005). A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical Report CBCL Paper #259/AI Memo #2005-036, Massachusetts Institute of Technology, Cambridge, MA.

Sur, M., Garraghty, P. E., and Roe, A. W. (1988). Experimentally induced visual projections into auditory

thalamus and cortex. *Science*, 242(4884):1437–1441.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, pages 1701–1708.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381:520–522.

Thorpe, S. J. (2002). Ultra-rapid scene categorization with a wave of spikes. In *Biologically Motivated Computer Vision*, pages 1–15.

Thorpe, S. J., Guyonneau, R., Guilbaud, N., Allegraud, J.-M., and VanRullen, R. (2004). Spikenet: Real-time visual processing with one spike per neuron. *Neurocomputing*, 58–60:857–864.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49:433–460.

Uchizawa, K., Douglas, R., and Maass, W. (2006). On the computational power of threshold circuits with sparse activity. *Neural Computation*, 18:2994–3008.

Uchizawa, K. and Takimoto, E. (2008). Exponential lower bound on the size of constant-depth threshold circuits with small energy complexity. *Theoretical Computer Science*, 407:474–487.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 18(11):1134–1142.

Valiant, L. G. (1994). *Circuits of the Mind*. Oxford University Press, New York, NY.

Valiant, L. G. (2000). A neuroidal architecture for cognitive computation. *J. ACM*, 47(5):854–882.

Valiant, L. G. (2005). Memorization and association on a realistic neural model. *Neural Computation*, 17(3):527–555.

Valiant, L. G. (2006). A quantitative theory of neural computation. *Biological Cybernetics*, 95(3):205–211.

Valiant, L. G. (2014). What must a global theory of cortex explain? *Current Opinion in Neurobiology*, 25:15–19.

Valiant, L. G. (2015). Personal communication.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience, New York.

Šíma, J. (2014). Energy complexity of recurrent neural networks. *Neural Computation*, 26(5):953–973.

Yang, G., Pan, F., and Gan, W.-B. (2009). Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462:920–924.