On the role of computational complexity theory in the study of brain function

Brendan Juba*

Abstract

We informally survey current work in the study of brain function with an eye for complexity-theoretic aspects. We then discuss in more detail why we expect computational complexity theory to play a larger, more explicit role in the future, examine the validity of such an approach, and attempt to outline how such a complexity-theoretic study might proceed.

Introduction

As you might be aware, Manuel Blum has recently been leading a project to develop a computer scientist's working definition of consciousness, called "CONSCSness," where CONSCS stands for CONceptualizing Strategizing Control Systems. Although Manuel has been interested in consciousness for a long time, at first glance it's a surprising move for him, given that his early career involved the development of computational complexity theory and the foundations of modern cryptography. As for myself, despite having worked with Manuel, Ryan Williams, and Matt Humphrey on CONSCS for a year or so, I tend to describe my interests as "complexity theory," and it is difficult to casually add, "I'm also working on consciousness." The statement always merits significantly deeper explanation: what business do I have or does any complexity theorist have, "working on" consciousness?

This piece is intended to answer that question, although I attempt to answer it in a way that pertains more broadly to studies of high-level aspects of brain function, freely replacing "consciousness" with "intelligence" or other such high-level concept. I hope to convince the reader of two things: one, that complexity theory is well suited to tackle problems arising in the study of brain function and two, that the sort of abstract theory that would arise from a computational complexity theoretic analysis of brain function would be highly desirable for solving certain basic problems, such as resolving whether or not attributes similar to consciousness or intelligence are present in other systems. This material bears some similarity in tone and content to portions of a talk Manuel gave at the 20th Computational Complexity Conference [2], but I hope specifically to explore the foundations of the program in a little more depth, and without assuming familiarity of much beyond the English language.

*bjuba@mit.edu. Carnegie Mellon University 5000 Forbes Ave. Pittsburgh, PA 15213. This work was supported in part by the National Science Foundation as part of the Aladdin Center (www.aladdin.cmu.edu) under grant CCR-0122581, and associated REU funding.

Background

Emerging views of the function of the brain

It has been generally accepted for a long time now that the basic functional unit of the brain is the neuron, and that the basic means for transferring information through the brain is via spikes in electric potential. Likewise, the basic “fire together, wire together” rule was put forth by Donald Hebb [7] over fifty years ago, and is now largely accepted in some form or another as the basic rule governing modification of the strengths of connections between neurons. Beyond this, it is well known from clinical data that certain structures or regions in the brain seem to come to serve certain functions.

It is worth noting that although some such large-scale patterns in the way brains are organized have been observed, there is still a remarkable diversity of structure from brain to brain. Also, brains are noted for their plasticity, their tendency to rewire and adapt to new inputs. Thus, although one could imagine, in principle, drawing a map of a given brain at one point in time by tracing the connections in these neural circuits, this map, this particular circuit is of limited significance to an overall theory of the brain’s function; it is merely one of numerous configurations that this particular brain would exhibit over its lifetime, to say nothing of the variety exhibited across different brains.

Now, out of the mountains of data that have been collected over the years, we are finally beginning to see some theoretical frameworks emerging to try to explain how high-level aspects of brain functions arise. In particular, basic frameworks for consciousness have been proposed by Francis Crick and Christof Koch [9], proposed by Gerald Edelman [3] (also with Giulio Tononi [18]), and proposed by others as well. Similarly, an informal characterization of intelligence and understanding was proposed by Jeff Hawkins [6] founded on high level frameworks for the interactions between regions of the cortex such as those proposed by Rajesh Rao and Dana Ballard [14] and Tai Sing Lee and David Mumford [10]. It is important to note that these frameworks attempt to explain brain function in terms of electrochemical neural activity in a comprehensible way, and are thus in contrast to proposals that the brain is too complex to be understood; that brain function arises in some significant way from quantum effects which would render some degree of knowledge about any brain, again, as “unknowable;” or that new laws of physics will need to be discovered to unravel the workings of the brain. Many believe that it is still too early to resort to such pessimistic positions—although the lack of concrete understanding of the brain’s workings might seem to support these proposals, a scientific program of study of the brain could not exist prior to the establishment of a guiding framework to supply hypotheses to be investigated.

Thus, these frameworks are an important step toward understanding how the brain functions, but all are still early revisions, which are likely to be refuted or rewritten in parts as data targeting these frameworks are collected in the coming years. One should note that all are tied in varying degrees to the architecture observed in the brain, though they do generally suggest what is functionally important about the architecture. While all of the proposals will probably require additional fleshing out in years to come, the explanatory powers of some of the more holistic accounts in particular leave much to be desired at present. The gaps in such accounts, particularly when some claim of *emergence* – the system having capabilities and properties that the individual components do not – is invoked, seem to fall neatly within the scope of computational complexity theory.

Relevant achievements and goals of computational complexity theory

We will be particularly interested in circuit complexity, the branch of computational complexity theory that studies how computational power varies with the number of elements used in a circuit, the "circuit size," or with the number of layers of elements used in a circuit, the "circuit depth," or both. Generally, a class of circuits is specified in terms of the types of elements involved (e.g., what gates are used and how many inputs they may have) and the size or depth of the circuit is specified with respect to the size of an input to the circuit. As an example, we will consider the class $\mathbf{AC}_0$: its members are families of circuits – one circuit for each number of inputs – for which the depths of the circuits are bounded by a constant, and the size of the circuit is bounded by any polynomial function of the number of inputs. These circuits operate on Boolean values and are comprised of the familiar AND, OR, and NOT gates with *unbounded fan-in*: that is, we permit an unlimited number of inputs to each AND and OR gate. Studies in circuit complexity generally take the form of establishing whether or not a particular class of circuits has members that can output the correct answer to any given input, thus solving a computational problem. I hasten to add that although we are most familiar with Boolean circuits, we are not necessarily restricting our attention to such circuits—in particular, there are several models of algebraic computation (applicable for circuits operating on real-valued inputs, for example) for which a sophisticated body of work exists; the ambitious reader is directed to Ben-Or's lower bound on the depth of Algebraic Computation Trees solving the element distinctness problem [1], as an example.

Demonstration that a class of circuits can solve some problem is often accomplished by constructing a circuit in the class that solves the problem in question, and is almost always straightforward in principle, at least. For example, it is easy to see that the problem of checking for a fixed substring can be solved by $\mathbf{AC}_0$: for any given string of interest, say 01100, in an input of length n we can compute a function such as

$$OR(AND(\neg x_1, x_2, x_3, \neg x_4, \neg x_5), \ldots AND(\neg x_{n-4}, x_{n-3}, x_{n-2}, \neg x_{n-1}, \neg x_n))$$

and it is clear that the resulting circuits are at most three gates deep and (if designed properly) never have size greater than $2n + 1$ gates. Meanwhile, arguments demonstrating that a particular class of circuits cannot solve some problem – a lower bound on the resources required for the problem – are notoriously difficult to conjure. Such lower bound results have been exhibited: an important result by Furst, Saxe, and Sipser, demonstrated that $\mathbf{AC}_0$ circuits cannot compute the parity of the input string [4]. Although such work was initially met with much optimism, it has not produced much stronger results than the aforementioned to date. Nevertheless, techniques for making lower bound arguments *do* exist, and these arguments comprise an important part of circuit complexity.

Now, observe that we could equivalently have defined $\mathbf{AC}_0$ to be circuits of merely OR and NOT gates, or just $NAND$ gates, and in any case, the class remains the same since the missing gates can be reconstructed with at worst a multiplicative constant increase in the depth and a polynomial increase in the size. These sorts of equivalences across different models of computation have been prominent from the earliest days of Computer Science, starting with the original Church-Turing thesis where it was observed that λ notation, Turing machines, recursive functions, and many other systems yielded equivalent classes of functions, which were then identified with the notion of effective computability. Since then, many computational problems have been identified as computationally equivalent to the problem of evaluating various computational models or *complete* for the class of problems solvable by those models. One of the strengths of the Theory of Computation is that we can work with whatever model is most convenient for our purposes, which merely happens to often be the Turing machine or familiar Boolean circuits.

Our larger goals have been inspired by the fact that computational complexity theory has been extremely successful at providing definitions of high-level concepts. One notable example is the definition of a function as "pseudorandom" by examining the probability, over random inputs to the function, of any detector circuit being fooled into claiming the function's output was random versus the probability of that detector correctly identifying a truly random string; if the size of the detector needed to achieve some degree of performance grows rapidly, we call the function pseudorandom. This definition has been made precise and, above all, has proved *useful* in demystifying pseudorandom number generators. We thus suspect that complexity theoretic ideas could be similarly useful in characterizing these high-level aspects of the function of the brain in more general and robust terms.

Current foundational work: the existing intersection

The first facet of an approach to studying the brain should be evident. We hope to identify the brain, or regions of the brain, with some class(es) of circuits. This idea is not new at all, and the consideration of circuits built using model neurons as gates dates back to the work of McCulloch and Pitts [11], at least. Once a class of circuits has been formally specified, we would then hope to be able to discuss precisely what sorts of functions could or could not be computed by such circuits, thereby bolstering or refuting claims about how the brain functions. Again, work has long been done based on our formal models of neurons along these lines, demonstrating the ability of a neural network to function as an associative memory [13], for example. In recent years, this challenge has been undertaken in a much wider scope by Leslie Valiant [20], who has been attempting to demonstrate the abilities of such simple models while seriously taking into account the sorts of parameters actually exhibited in the brain. This sort of work is precisely what is needed to flesh out claims of emergence in theories of brain function, since it considers precisely whether or not it is possible for the model system to exhibit the claimed properties.

I should also discuss the recent work on modeling the feedforward path of the visual cortex, carried out independently by Thomas Serre et al. [15], and Simon Thorpe et al. [16, 17], not only for its significance to the project at hand, but also since it 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 highly 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 [16] 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. [5] 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. [15] 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, tests that have not yet been performed for the model of Thorpe et al. While I have no doubt that the community will soon sort out which of these models, if either, is likely to be correct, I claim that the lesson we should take away from this incident is that 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. From our prior discussion of equivalences across models of computation, I hope to have made it clear that often a given class of problems will have many equivalent formulations. In other words, it's frequently possible for one model to simulate another and vice-versa, and from a functional standpoint, it does not matter which model we use. I should remark that such an idea is not entirely foreign to biologists and neuroscientists; the related notion of different circuits that exhibit equivalent behaviors (which we certainly would expect to encounter in the cortex) is discussed at length by Edelman [3], who calls such systems *degenerate*. The difference in Edelman's notion of degeneracy is that he is considering equivalent structures in a fixed model, whereas here we are concerned with the possibility that two different models 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 will generally not contribute to our understanding of its function. I should note now that it is not immediately clear to me whether or not the models of Serre et al. and Thorpe et al. are computationally equivalent; the point is only that the functionality essentially never identifies a unique model, to say nothing of the underlying physiology that (in a sense) implements the model. Having made this distinction, I would like to make explicit that the object of study in computational complexity theory *is* the relationship between such a bounded computational model and its functional capabilities.

About the approach

There are two distinct programs of study in which I foresee computational complexity theory playing two rather different roles. The first is in continuing the low-level studies of Serre et al. [15], Thorpe et al. [16, 17], and Valiant [20], where the objective is to understand which functions the brain is computing. In continuing this study, I propose that treating highly plastic regions of the brain as an arbitrary member of an appropriate class of circuits and performing a complexity-theoretic analysis of this class may be helpful in understanding its functional capabilities. The second program of study aims to explain how the activity of the brain gives rise to familiar high-level properties such as intelligence and consciousness, following work by Crick and Koch [9], Edelman and Tononi [18, 3], Rao and Ballard [14], and Lee and Mumford [10]. Manuel's CONSCS project [2] is such a study; the role of computational complexity theory in service of this program is largely to provide a vocabulary for defining properties that we would identify with these phenomena in a sufficiently general terms that we can use them across domains. In addition to outlining how complexity theory could facilitate these programs, I will also discuss why the ability to transfer theories of high-level brain function across domains is significant and how such a study is justified.

The role of complexity theory in the low-level study of brain function

Turning to the future now, as we move further from the regions of the cortex receiving inputs, we also enter the more plastic regions of the cortex, where we expect to encounter a wide variety of functionality. In discussing what these circuits do, it is probably most appropriate to talk in terms of a complexity class: since in such regions the brain is known or believed to frequently "rewire" itself while the number of neurons remains roughly bounded across time and different subjects, the class of functions that can be expressed using at most the number of neurons in these regions seems to be an appropriate characterization of what these regions are capable of, in principle at least. Certainly, given the sheer diversity of the brain's functionality, we know that brains have computed *some* vast class of functions, so unraveling the functionality of the brain is going to require understanding some such class of functions, with the class yielded by bounding the neural circuit size being a natural candidate. In any event, this is no small feat.

Fortunately, Mountcastle's common cortical algorithm hypothesis [12], which states that the underlying model elements behave essentially the same across the cortex, should permit us to at least understand how the elements of these circuits behave by examining them in, say, the visual areas; we expect that the only distinction across regions of the cortex should be the parameters such as the size, depth, connectivity, etc. of the circuits. Thus, a complexity-theoretic analysis of these regions should be feasible. By contrast, it is not clear at all how the current techniques will manage to unravel the workings of these regions, since the variety of different behaviors we exhibit is daunting, and the relationship between these behaviors and the recordings we would obtain in such regions is murkier: in the regions of the brain more distant from the points of entry for its inputs, our ability to study its response to a controlled stimulus (such as it is) will only grow more limited, especially since we lack the ability to track which other neurons a given neuron is signaling. Ultimately, I expect that new techniques will need to be developed to try to better understand how functionality in the brain is controlled, but I anticipate that a nothing less than a complexity-theoretic analysis will capture all of the functionality of the brain. In any case, It will *certainly* be essential to transferring what we learn about the brain to and from studies in other settings.

The necessity of abstract and generalized theories

In the study of the function of the brain, following Manuel [2], we 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 study of the physical brain: in the physical study, certain processes would be observed which could be verified to have the necessary properties to be called "consciousness" or "intelligence" in the abstract theory. These theories are desirable because they are *necessary* for the study of our high-level features in other domains.

Specifically, if we wished to decide whether or not an ant colony or a robot is conscious or intelligent, it would obviously be unsatisfying to conclude that since neither of these two objects use a primate's brain (where most of the physical studies have been carried out), neither one could be conscious or intelligent. At the other extreme, although Turing's test [19] in which a judge converses with a human and a machine, attempting to tell which is which is a clever attempt at distinguishing intelligence without knowing what "intelligence" is, the test has some serious shortcomings, minimally including 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 implementations 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. While the theory should also make it possible in principle to demonstrate conclusively that a system is *not* intelligent or conscious, I would hesitate to claim that demonstrating the absence of these properties would necessarily be so simple. It is entirely possible that arguing such a claim would necessitate demonstrating that the system in question is incapable of computing some necessary functions using some lower-bound argument.

Regardless of the difficulties that might arise in demonstrating that a system cannot be conscious, intelligent, etc., merely being able to demonstrate that some systems do have such properties should be extremely beneficial to AI. Although it is entirely possible that an AI researcher will implement a system that satisfies all of our conditions for being intelligent without a clear idea of what those conditions are, it should be clear that this sort of blind success is doubtful. By contrast, once the requirements for machine intelligence are clearly stated, it is likely that researchers will be able to solve the problem of building an intelligent machine—after all, this problem has been solved in the workings of the brain, so it is reasonable to expect that the problem is at least computationally tractable, if not simple.

Evasion of philosophical debates: the validity of the program

It is presently worth examining the sort of philosophical side step that has already been employed in the scientific study of consciousness. To avoid being bogged down by debates over physicalism versus dualism – for example, whether or not consciousness is due to some undetectable "soul stuff" as proposed by Descartes – and the like, Francis Crick and Christof Koch have emphasized that they are only studying the Neural Correlates of Consciousness (NCC). They have stated the largely non-contentious hypothesis that mental states should be correlated with certain neural states. This maneuver has been employed by many others as well in founding scientific studies of consciousness, and the underlying hypothesis is well-supported by a variety of data. The interested reader is referred to Koch's book, *The Quest for Consciousness* [9].

Rather than discuss the details of the hypothesis and the supporting evidence, we'll consider the implications of taking this side step. Certainly, from a physicalist's point of view, nothing has been lost, since the NCC are equated with consciousness—a physicalist would expect that a property such as consciousness is equivalent to some properties of neural activity, where the aim of the programs is clarification of this relationship. Even a dualist, though, would have to honestly admit that studying the NCC would have some practical value. For example, once the NCC have been identified, as Koch points out [9], it would be conceivable to build a device to measure consciousness or to develop more effective anesthetics. What is most relevant about this maneuver, though, is that by limiting their presumed scope to the measurable aspects of consciousness, they study precisely the aspects of the problem that are tractable by science; that is, from a scientific point of view, no part of the problem has been put off-limits, and the presented part of the problem is guaranteed to be within the reach of science. Of course, studying the NCC is unlikely to resolve the philosophical debate entirely (at best, one would be able to invoke Occham's Razor), but our point is that one need not resolve the philosophical concerns to obtain something useful.

I have dwelt on this point for two reasons: first, a computer scientist's study of consciousness would be a study of these NCC, and second, our abstracted study will need to take similar side-steps. We can only claim, for example, that a system satisfying our properties will simulate the

NCC. By defining in what sense we "simulate," we will be taking another such side step away from the philosophers' notion of consciousness.

Outline of a complexity-theoretic study of high-level brain function

As suggested above, 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." We begin by viewing the brain as a formal system, as circuits from a particular class, where the class would be identified from studies such as those currently being carried out in the previously mentioned work of Serre et al. [15], Thorpe et al. [16, 17], and Valiant [20]. We continue by asking what properties of those circuits characterize our concepts like 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, let's try to give a first attempt at a complexity-theoretic definition of "intelligence." (See Manuel's talk [2] for a similar initial attempt at defining "consciousness.") We will follow Hawkins' proposal [6] that intelligence is the ability to make predictions about one's environment which was supported by the work of Rao and Ballard [14], Lee and Mumford [10], and others who suggested that the feedback connections in the cortex serve to carry predictions about future inputs. We 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. We will consider environments to be represented by sequences of input strings, $x_1, x_2, \dots, x_t, \dots$, although we will later have cause to consider *classes* of environments, which are merely sets of possible input string sequences. We will consider our candidate intelligent machine M to have some internal state s_t, which in the language of Turing machines would be the contents of its worktapes 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 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 strictly better than chance. Clearly, 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 clearly 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 $\mathcal{C}$.[1]

[1] The definition obtained in this way is in contrast to that proposed by Hutter [8], who only considered a goal-oriented setting and focused on the entity's behavior, which was assumed to be in pursuit of maximizing reward.

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 intelligent with respect to a constant environment, but no machine is intelligent with respect to 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 nature. 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 timestep would be computed from portions of its state in the previous timestep 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. As part of defining intelligence, we would like in general to characterize these environments more precisely; machines intelligent with respect to 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 model that would satisfy it to have some computational power.

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." 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; that is, we would like to separate the specifications of the NCC, 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, understanding, etc. Manuel has used a term like "CONSCS" to emphasize that we are only seeking a working definition for our formal systems: we are only seeking a characterization of consciousness (properly, the NCC) when the brain is viewed as a formal system. In this way, again, because we are only claiming to study a formal system, the only contentious point is whether or not results about our formal system will be relevant.

We do expect such results 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 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 general complexity theoretic results 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.

Conclusion

As our understanding of the low-level function of the brain has improved to the point where Serre et al. [15] and Thorpe et al. [16, 17] can propose plausible quantitative models, scientific communities have begun to seriously explore topics such as consciousness. Frameworks attempting to explain how the workings of the brain give rise to consciousness have been proposed by Edelman [3], by Crick and Koch [9], and by others, and an attempt at explaining intelligence has been offered by Hawkins [6] based on the frameworks of Rao and Ballard [14], Lee and Mumford [10], and others. We can expect that in the coming years, these theoretical frameworks will develop into a reasonable account of how certain high-level attributes arise from the low-level function of the brain.

While it's evident that researchers in other fields are making progress toward comprehending the function of the brain, I hope I have convinced you that we, as complexity theorists, are justified in working on brain function and moreover, that the present is an appropriate time for complexity theorists to dive into the study of brain function. In many cases, hard results in complexity would flesh out, bolster, or refute claims used in theories of the function of the brain. Moreover, computational complexity theory is well-positioned to provide exactly the sort of precise yet "machine-independent" theories of brain function that will be required for certain applications, as we have had background in defining other high-level concepts, such as randomness, in this way. These similar sorts of programs have also been particularly successful for complexity theorists: we have provided rigorous, useful definitions. Although I am confident that the researchers currently working out the theory of these aspects of brain function could, in time, develop the necessary tools to decipher the workings of the brain, I am likewise certain that the process will be expedited if complexity theorists lend a hand.

Acknowledgements

Thanks to Erick Chastain for much helpful feedback and input, and to Leslie Valiant for a helpful and informative discussion. Thanks also to Manuel Blum, Ryan Williams, and Matt Humphrey for the many interesting and influential discussions while working on the CONSCS project.

References

[1] Michael Ben-Or. Lower bounds for algebraic computation trees. In *STOC '83: Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 80–86, New York, NY, 1983. ACM Press.

[2] Manuel Blum, Ryan Williams, Brendan Juba, and Matt Humphrey. Toward a high-level definition of consciousness. In *20th IEEE Computational Complexity Conference*, San Jose, CA, 2005. Invited Talk. `http://www.cs.cmu.edu/~mblum/research/pdf/CONSCS8.ppt`.

[3] Gerald M. Edelman. *Wider than the Sky: the Phenomenal Gift of Consciousness.* Yale University Press, New Haven, CT, 2004.

[4] Merrick L. Furst, James B. Saxe, and Michael Sipser. Parity, circuits and the polynomial hierarchy. *Math. Systems Theory*, 17:13–27, 1984.

[5] Rudy Guyonneau, Rufin VanRullen, and Simon J. Thorpe. Neurons tune to the earliest spikes through STDP. *Neural Computation*, 17:859–879, 2005.

[6] Jeff Hawkins and Sandra Blakeslee. *On Intelligence.* Times Books, New York, NY, 2004.

[7] Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory.* Wiley, New York, NY, 1949.

[8] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2004.

[9] Christof Koch. *The Quest for Consciousness: a Neurobiological Approach.* Roberts & Company Publishers, Englewood, CO, 2004.

[10] Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America*, 20(7):1434–1448, 2003.

[11] Warren S. McCulloch and Walter Pitts. A logical calculus of ideas immanent in neural activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

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

[13] Günther Palm. *Neural Assemblies: an Alternative Approach to Artificial Intelligence.* Springer-Verlag, Berlin, 1982.

[14] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999.

[15] Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio. 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, 2005.

[16] Simon J. Thorpe. Ultra-rapid scene categorization with a wave of spikes. In *Biologically Motivated Computer Vision*, pages 1–15, 2002.

[17] Simon J. Thorpe, Rudy Guyonneau, Nicolas Guilbaud, Jong-Mo Allegraud, and Rufin Van-Rullen. Spikenet: Real-time visual processing with one spike per neuron. *Neurocomputing*, 58–60:857–864, 2004.

[18] Giulio Tononi and Gerald M. Edelman. Consciousness and complexity. *Science*, 282(5395):1846–1851, 1998.

[19] Alan M. Turing. Computing machinery and intelligence. *Mind*, 49:433–460, 1950.

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