A conversation with Professor Shaul Druckmann, September 5, 2019

Participants

- Professor Shaul Druckmann - Assistant Professor of Neurobiology and of Psychiatry and Behavioral Sciences, Stanford University
- Joseph Carlsmith - Research Analyst, Open Philanthropy

Note: These notes were compiled by Open Philanthropy and give an overview of the major points made by Prof. Druckmann.

Summary

Open Philanthropy spoke with Prof. Shaul Druckmann of Stanford University as part of its investigation of what we can learn from the brain about the computational power ("compute") sufficient to match human-level task performance. The conversation focused on the compute necessary to model different processes in the brain.

Categorization of neural processes

Mr. Carlsmith suggested the following categorization of signaling processes in the brain:

- canonical neuron signaling (e.g., mappings from synaptic inputs to firing decisions, and from pre-synaptic spiking to immediate impacts on the post-synaptic neuron),
- learning (e.g., experience-dependent changes to neurons and synapses over time),
- alternative signaling mechanisms (e.g. glia, neuromodulation, ephaptic effects, and others),
- other unknowns.

This sounded like a reasonable classification to Prof. Druckmann.

Setting aside plasticity, most people assume that modeling the immediate impact of a pre-synaptic spike on the post-synaptic neuron is fairly simple. Specifically, you can use a single synaptic weight, which reflects the size of the impact of a spike through that synapse on the post-synaptic membrane potential.

Experimental barriers

Here are a few examples of experimental barriers faced by neuroscientists, which make it difficult to determine the level of biophysical detail models need to include in order to capture the dynamics relevant to the brain's task-performance *in vivo*.

Uncertainty about inputs and outputs

It is very difficult to tell what spatio-temporal patterns of inputs are actually arriving at a neuron's synapses *in vivo*. You can use imaging techniques, but this is very messy.

Similarly, it is very difficult to tell how a neuron responds to arbitrary patterns of synaptic input. You can stimulate a pre-synaptic neuron and observe the response, but you can't stimulate *all* pre-synaptic neurons in different combinations. And you can only patch-clamp one dendrite while also patch-clamping the soma (and this already requires world-class skill).

Ion channel densities

Technology for measuring the properties relevant to detailed biophysical modeling has improved very little in the past 20 years (though technology for recording the spiking of neurons has improved a huge amount). Neurons can have a few dozen of some 200-300 types of ions channels, which are strongly non-linear, with large effects, and which are spread out across the neuron. These cannot be modeled based on recordings of neuron spiking activity alone, and staining neurons for these ion channels is very difficult.

In vitro experiments

Experiments involving current injection into the soma are typically done *in vitro*, because it is easier to patch-clamp neurons *in vitro*.

Limitations of animal models

The tasks that neuroscientists tend to study in model animals are very simple. Many, for example, are some variant on a two-alternative forced choice task (e.g., teaching an animal to act differently, depending on which of two stimuli it receives). This task is extremely easy to model, both with a small number of highly simplified neurons, and with models that do not look like neurons at all. In this sense, tasks like these provide very little evidence about what level of modeling detail is necessary for reproducing more interesting behavior.

Attempts at simplification

We can distinguish between two approaches to the brain's biophysical complexity. One camp argues: "let's not assume we need to include a given type of biophysical complexity in our models, until doing so becomes necessary." The other argues: "If this complexity were in fact important, we would not currently be able to tell." Prof. Druckmann tends to be in this latter camp, though he thinks that the former is a fair and practical approach.

Simplifying complex models

Some neuroscientists have explored the possibility that detailed biophysical models can be reduced to simpler neuron models or equivalent deep neural networks. Biophysical models involve a very large number of differential equations, so we know that there is some input regime in which it would be impossible to capture what this model does without a giant neural network, since there are a large number of non-linear operations that are difficult to approximate. But we don't know how much of that regime is actually sampled in the neuron's normal life.

What's more, many dendritic non-linearities contribute more strongly when triggered by synaptic inputs arriving at similar times to similar dendritic locations ("clustering"), and there is evidence that such clustering occurs *in vivo*. In this sense, a random input regime is unrepresentative, more weakly non-linear than it should be and therefore may be particularly easy to model.

Prof. Druckmann does not think that appeals to the manageable compute burdens of modeling of dendrites as comparatively small multi-layer neural networks (for example, with each dendritic sub-unit performing its own non-linearity on a subset synaptic inputs) definitively address the possibility that modeling dendritic non-linearities requires very large amounts of compute. Small multi-layer network models are really just a guess about what's required to capture the neuron's response to realistic inputs.

For example, in a recent unpublished paper, David Beniaguev, Idan Segev, and Michael London found that adding NMDA currents to the detailed model increased the size of the neural network required to replicate its outputs to seven layers (the long time-constant of NMDA receptors increases the complexity of the neuron's input-output transformation). Adding in other neuron features could require many more layers than this. 10 layers might be manageable, but 500 is a pain, and the true number is not known.

Neuron modeling competition

One neuron modeling competition proceeded by assuming that dendritic inputs are randomly distributed, and that dendrites just integrate inputs linearly -- assumptions used to create a pattern of current to be injected into the soma of the neurons whose spikes were recorded. If these assumptions are true, then there is good reason to think that fairly simple models are adequate. However, these assumptions are very friendly to the possibility of non-detailed modeling. The point of complex models is to capture the possibly non-linear dendritic dynamics that determine what current goes into the soma: after that point, modeling is much easier. And we don't know to what extent non-random inputs trigger these dendritic dynamics.

There were also a few other aspects of this neuron modeling competition that were not optimal. For example, it was fairly easy to game the function used to evaluate the models.

More complex models

Prof. Druckmann does not think it obvious that the kind of multi-compartmental biophysical models neuroscientists generally use are adequate to capture what a neuron does, as these models, too, involve a huge amount of simplification. Calcium dynamics are the most egregious example. Real neurons clearly do things with calcium, which moves around the cell in a manner that has consequences for e.g. calcium-dependent ion channels. Most biophysical models, however, simplify this a lot, and in general, they treat ions just as concentrations affected by currents.

There are also probably other things that matter that these models leave out, but that we aren't aware of.

Examples of the relevance of low-level details

There are some edge cases where specialized biophysical dynamics are known to be important -- for example, in the Calyx of Held, a synapse that allows for precise timing in the auditory cortex. But it's not clear how generic such cases are.

Hypotheses about the brain's architecture

In the 1980s and 1990s, many neuroscientists thought of the brain's computations in terms of sensory-motor transformations. However, it is unclear that you actually need a huge brain for this: flies, for example, can see and walk around fairly well, but their brains are tiny.

Prof. Druckmann thinks that the need for large brains grew out of the need to create more sophisticated models of the world. It's easy to underestimate the complexity of the world-models humans work with, which involve millions of objects with arbitrary dynamics in space and time. The brain is keeping a running tally of many things in the world, without your conscious knowledge. For example, if one of your roommates drinks a lot of the milk while you're asleep, the next morning you will lift up the container too forcefully.

Matching the brain's modeling in this respect is beyond the current reach of deep neural networks, and may require greater understanding of the logic behind the brain's architecture. For example, the brain is much more recurrent than deep neural networks. It would be nice to know what role this recurrence is playing.

There is also another school of thought that hypothesizes that the brain's complexity emerges centrally in order to facilitate efficient learning (much more efficient than current reinforcement learning systems). Neuroscientists aren't currently in a position to tell if this is true.

The possibility of simplification

There is no evidence that replicating the brain's task-performance strictly requires replicating its biophysical dynamics, and it's unclear what such evidence would even look like. After all, we can model almost any input-output transformation using a sufficiently complex neural network.

The argument would need to be that the brain's methods are much more efficient. But in order to make and evaluate this type of argument, we need to understand the brain's computation much better than we do, and we are very far away from being able to evaluate such claims using actual experiments.

Prof. Druckmann would be extremely surprised if future working models of human intelligence incorporate large amounts of biophysical detail (e.g., molecular dynamics). He is confident that the type of non-linearities generated by real biophysics can be more efficiently emulated in different ways in a model. Therefore, these models will look more like giant networks of simple artificial neurons than giant networks of Hodgkin-Huxley models.

The route forward

The question is: what is the best route to achieving the type of understanding necessary to know which aspects of the brain's architecture can be bypassed? In this respect, Prof. Druckmann is sympathetic to approaches that proceed from the bottom up -- e.g., by attempting to understand how the brain works in detail, as a means to understanding how to simplify it.

This sympathy emerges centrally out of pessimism about the progress made by alternative approaches. Prof. Druckmann feels that neuroscience hasn't yet managed to make a lot of progress on the general question of the detailed logic of building a brain, or a cortical area. A conceptual understanding of computation in strongly recurrent, broadly distributed networks is still in its infancy. This is true both of conceptual/formal approaches to the problem, and of approaches in engineering and robotics.

In this sense, Prof. Druckmann believes that work in the 80s by David Marr was overly optimistic about how easy biology would be (though he admires Marr's work regardless). Outside of early sensory systems, we have made very little progress in understanding the brain at a normative or algorithmic level. This makes Marr's research program very difficult. Accordingly, attempting to proceed via the implementation level is the option that remains.

Constraints faced by the brain

Once we figure out how the brain works, though, there's no reason to think you couldn't implement the same computations in a different way. Any time you have a dynamical system, and you're only interested in one slice of that system's behavior, it's unlikely that replicating the whole system is the only way to replicate that slice.

This seems especially unlikely given that the brain was designed to meet many constraints that would not apply to human engineers. For example, artificial systems probably won't need to include the same repertoire of ion channels, as non-linearities can be encoded in different ways. Similarly, the brain has to learn at a certain rate, has to be specifiable genetically, has to fit through the birth canal and so forth.

Sources of neuroscientific opinion

Prof. Druckmann believes that at our current conceptual understanding of neural computation, many statements in neuroscience to the effect that "we can reduce X to Y" are based mostly on personal opinion, sometimes influenced in part by what current technology allows us to do, rather than in well-justified, first-principles reasoning. The

history of neuroscience sometimes seems like a process in which even though some process or level of detail is important, if it is very difficult to understand it, the community often shifts away from that level, and moves on to another level.

The decrease in investment of the general neuroscience community in single neuron biophysics is one example. The field has basically given up on detailed biophysical modeling. In the 1990s, there were many papers in top journals on the topic, but now there are almost none. Prof. Druckmann expects that the large majority of people who do not work in early sensory systems would say that detailed biophysical modeling is unnecessary for understanding the brain's computation.

A charitable rationale for this appeals to the possibility that more detailed dynamics average out at the level of neuron populations. But Prof. Druckmann does not think that there are real results that definitively suggest this. Rather, he thinks that people don't do detailed modeling because these models are ill-constrained at the current level of data that can be collected and it would require major investment to get the relevant data.

Prof. Druckmann also does not think that neuroscientists have much of an incentive to settle the debate about what level of modeling detail is adequate to match the human brain's computational complexity. Most do not expect the timescale of their career to include a working copy of the human brain, the work they are currently doing does not require detailed modeling, and it's very hard to have informed views about what will ultimately be necessary.