A lightly-edited email conversation between Adam Marblestone and Holden Karnofsky, April 13, 2016

Participants

- Adam Marblestone – Director of Scientific Architecting, Massachusetts Institute of Technology Synthetic Neurobiology Group (scientific advisor to the Open Philanthropy Project)
- Holden Karnofsky – Executive Director, Open Philanthropy Project

Note: this set of notes was compiled by the Open Philanthropy Project.

Summary

Holden Karnofsky wrote to Adam Marblestone about the possibility that the brain uses a relatively small number of basic algorithmic approaches or learning rules to do most important cognitive tasks.

Lightly-edited email from Holden Karnofsky, 4/13/2016

Hi Adam,

I hope you're doing well. I'm working on a blog post about the possibility of transformative artificial intelligence, and one of my claims is that it seems *possible* (not necessarily true, but possible) that the human brain uses a relatively small number of basic algorithmic approaches in order to do most important cognitive tasks. My understanding, from other conversations, is that:

- The current state of neuroscience is consistent with the "small # of algorithms" idea, i.e., it's an open question.
- This "small # of algorithms" idea was originally floated in the 1970s but fell out of favor, and it was first resurrected by Jeff Hawkins's suggestion of cortical similarity (though one could reject cortical similarity while still endorsing "small # of algorithms").
- The idea has since gained somewhat wider acceptance, but cortical similarity has limited mainstream support, and not everyone who endorses "small # of algorithms" endorses cortical similarity.

Does that sound right? If so, would you be able to point me to a review paper or other credible source that can demonstrate that the "small # of algorithms" idea is a live possibility/open question?

Any thoughts welcome.

Thanks!

Best,
Holden

Lightly-edited email from Adam Marblestone 4/13/2016

[Note: quotes from the earlier Holden Karnofsky email are indented below here.]

Hey Holden!

Glad you're thinking about these questions. My answer is somewhat complicated, so I hope you can bear with me in the below, and please feel free to follow up with more thoughts or questions of course. See inline.

> Hi Adam,
>
> I hope you're doing well. I'm working on a blog post about the possibility of transformative artificial intelligence, and one of my claims is that it seems *possible* (not necessarily true, but possible) that the human brain uses a relatively small number of basic algorithmic approaches in order to do most of the important stuff it does.

It depends on what you mean by "algorithmic approaches". I would distinguish "learning rules" from "computations" in this context. Learning rules are how the brain would, in a way that could be dependent on the input data and on various forms of pre-structuring of the circuits, transform itself from a relatively random wiring to a wiring that can actually do useful computation. In an artificial neural net, for example, the learning rule is usually backpropagation everywhere, but inside the network, this can give rise to a huge number of different computations. So you could imagine that, say, the cortex could have a "common learning rule" in the sense that the circuits would be shaped by data and by training signals in similar ways throughout the cortex, or, more strongly, it could have a "common computation", in the sense that, in real time, all the areas are running the same algorithm, i.e., processing their data in the same way in real time. In the case of Hawkins, there is a kind of mixing of the two, but in general I would distinguish these.

Yet another option is that there is a common communication interface throughout the cortex, but that the circuitry, learning rules and/or computations are actually quite specialized in each cortical area. So "common computation", "common learning rule" and "common communication interface" are three options for the cortex, in increasing order of plausibility in my personal opinion (see below).

It also depends what you mean by "small". Some people argue that the cortex has the same learning rule all over. But there are many other brain areas besides the cortex. For example, the basal ganglia seems to be heavily involved in decision making and reinforcement learning, the thalamus seems to be heavily involved in

long-range information routing, the hippocampus seems to be involved in memory and also in spatial navigation or perhaps more generally in the construction of a simulation of the environment. Hawkins argues that the hippocampus is "just" the "top layer" of the cortical hierarchy, but there is not consensus on this point. So even if the cortex is uniform in terms of learning rule, then still a) that learning rule could produce many local computations in different areas, and b) there are still all the other areas besides cortex.

Nevertheless, even if there are, say, 20 or 30 fundamental algorithmic processes in the brain that underly intelligence, I would still count that as "small" in the sense that it is possible that computational neuroscience could actually seek to understand all of them and how they interact, and use that to inform the development of AI.

My understanding, from other conversations, is that:

- The current state of neuroscience is consistent with the "small # of algorithms" idea, i.e., it's an open question.

I would agree. It is an open question. It is definitely not a crazy idea.

Gary Marcus, Tom Dean and I wrote an essay a year or two back, where we tried to call into question the issue of cortical uniformity. The very fact that we had to write that essay shows you that at least some significant community considers a "small number of algorithms" view quite plausible.

Here is the essay (peer-reviewed):

http://cs.brown.edu/people/tld/publications/archive/MarcusandMarblestoneandDeanSCIENCE-14.pdf

And here is a non-peer-reviewed FAQ for it, which goes into more detail:

http://arxiv.org/abs/1410.8826

*One thing to look for in there is all the mechanisms biology has available to tinker with and tweak circuitry, that might be invisible to us right now at the level of precision and scale brain mapping that we have (i.e., the level where the cortex looks pretty uniform). In general, biology exhibits a lot of very specialized structure, and evolution can "duplicate and diverge" that structure so as to tinker with it. Consider the gene expression programs that give rise to hand vs. foot, for instance. They are related, but meaningfully different. I bet the cortex is like that too, and the question is: what does the diversity do. Below I suggest that it may specify different cost functions for learning, but who knows.)

Gary Marcus gave informative talks about this work at several venues:

https://archive.org/details/Redwood_Center_2014_09_19_Gary_Marcus
https://www.msri.org/workshops/796/schedules/20448
http://digitalops.sandia.gov/Mediasite/Play/9d41b9923e834170bfe1dc7cf52e353f1d
https://www.youtube.com/watch?v=Ux7vPuTguYQ

I would say that those talks from Gary kind of present a bit of the for, and a bit of the against, as far as cortical uniformity.

Now, I think the table in the FAQ document is kind of in the framework of "what are the computations", whereas a better question would perhaps be "what kinds of learning rules are there"? But even with our attempt to emphasize diversity of computations, we still end up with a finite list. The details are probably wrong, but my current hunch or hope, actually, is that there will be a way to functionally decompose brain computation into some reasonable-sized list like this, which could serve as a kind of architectural sketch of the brain.

I am currently working with some collaborators on another essay that takes a different view. There, we would have at least one very powerful learning algorithm, essentially a biologically plausible equivalent of backpropagation for neural nets (various such have been proposed in the literature), which allows multi-layer credit assignment in a deep network. It would be a work-horse, in terms of generating, based on learning, the actual computations that end up occurring in the final circuitry.

If we have such a powerful learning algorithm, then the question is: how it would be put to use to generate a diversity of computations and processes in the brain, much as machine learning puts backpropagation to use to do a variety of things. We try to frame that in terms of the idea of "cost functions", arguing that the brain may internally generate a wide diversity of cost functions for training its networks, but that perhaps the training itself is done using a fairly generic and powerful kind of learning mechanism.

- This “small # of algorithms" idea was originally floated in the 1970s but fell out of favor, and it was first resurrected by Jeff Hawkins's suggestion of cortical similarity (though one could reject cortical similarity while still endorsing "small # of algorithms").

I'm not sure of the exact history. But that sounds basically plausible. Recall that everything in these fields has been under debate for decades and that things come into and out of fashion. Also recall that there are, as of today, exactly zero detailed connectomic maps of cortex across areas that could "ground truth" these issues to some extent.

When was Minsky's "Society of Mind" book published? You could perhaps view that as kind of the opposite of a "uniform statistical learning rule" view. Recall that the response to Minksy and Papert's earlier book on Perceptrons also kind of put on hold the field of neural nets for a while in AI, viewing them as too simple to learn to do powerful cognitive computations. What we now know is that even simple learning rules can, with the right architecture and training data, give rise to extremely complex computation. But even Society of Mind might be described as having a small number of essential design principles or architectural elements at work. In fact I think it is consistent with a view in which there are a small number of key developmental and learning principles, but these interact and bootstrap off of one another in a complicated, and genetically quite precisely orchestrated, way.

Evolutionary psychology is also a relevant piece of intellectual history here, where the idea of a number of evolved modules selected to perform very specific functions is perhaps contrary to the idea of a small number of algorithms, but not necessarily so I think.

- The idea has since gained somewhat wider acceptance, but cortical similarity has limited mainstream support, and not everyone who endorses "small # of algorithms" endorses cortical similarity.

It is true that not everyone who endorses small # algorithms endorses cortical similarity. I might be in that camp, for instance, insomuch as I have any defined opinion at this early stage of brain research in which we still lack ground truth knowledge of mesoscale neural circuitry.

Hawkins, who has been a vocal proponent of cortical uniformity, is perhaps viewed as outside of mainstream academia based on his career history, but he does attend scientific meetings and publish papers, such as this one http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000532 - which has been very influential, as well as more recently - http://journal.frontiersin.org/article/10.3389/fncir.2016.00023/full - which makes a number of specific predictions that would be testable.

Certainly far from being agreed on or proven right, and perhaps even likely to be wrong at some level, as is arguably the nature of much theory at this stage of neuroscience. How many truly ambitious papers in computational neuroscience today could not be subject to those same risks or criticisms?

> Does that sound right? If so, would you be able to point me to a review paper or other "legitimate/mainstream/credible" source that can demonstrate that the "small # of algorithms" idea is a live possibility/open question?

See above. It is definitely open. I think it is one of the most important open questions in neuroscience. But "small" might be 20 things, not 1 thing, in my view.

All Open Philanthropy Project conversations are available at
http://www.openphilanthropy.org/research/conversations