# Appendix: Is there a *simple* algorithm for intelligence?

In this book, we've focused on the nuts and bolts of neural networks: how they work, and how they can be used to solve pattern recognition problems. This is material with many immediate practical applications. But, of course, one reason for interest in neural nets is the hope that one day they will go far beyond such basic pattern recognition problems. Perhaps they, or some other approach based on digital computers, will eventually be used to build thinking machines, machines that match or surpass human intelligence? This notion far exceeds the material discussed in the book - or what anyone in the world knows how to do. But it's fun to speculate.

There has been much debate about whether it's even *possible* for computers to match human intelligence. I'm not going to engage with that question. Despite ongoing dispute, I believe it's not in serious doubt that an intelligent computer is possible - although it may be extremely complicated, and perhaps far beyond current technology - and current naysayers will one day seem much like the vitalists.

Rather, the question I explore here is whether there is a *simple* set of principles which can be used to explain intelligence? In particular, and more concretely, is there a *simple algorithm for intelligence*?

The idea that there is a truly simple algorithm for intelligence is a bold idea. It perhaps sounds too optimistic to be true. Many people have a strong intuitive sense that intelligence has considerable irreducible complexity. They're so impressed by the amazing variety and flexibility of human thought that they conclude that a simple algorithm for intelligence must be impossible. Despite this intuition, I don't think it's wise to rush to judgement. The history of science is filled with instances where a phenomenon initially appeared extremely complex, but was later explained by some simple but powerful set of ideas.

Consider, for example, the early days of astronomy. Humans have known since ancient times that there is a menagerie of objects in the sky: the sun, the moon, the planets, the comets, and the stars. These objects behave in very different ways - stars move in a stately, regular way across the sky, for example, while comets appear as if out of nowhere, streak across the sky, and then disappear. In the 16th century only a foolish optimist could have imagined that all these objects' motions could be explained by a





## Resources

simple set of principles. But in the 17th century Newton formulated his theory of universal gravitation, which not only explained all these motions, but also explained terrestrial phenomena such as the tides and the behaviour of Earth-bound projecticles. The 16th century's foolish optimist seems in retrospect like a pessimist, asking for too little.

Of course, science contains many more such examples. Consider the myriad chemical substances making up our world, so beautifully explained by Mendeleev's periodic table, which is, in turn, explained by a few simple rules which may be obtained from quantum mechanics. Or the puzzle of how there is so much complexity and diversity in the biological world, whose origin turns out to lie in the principle of evolution by natural selection. These and many other examples suggest that it would not be wise to rule out a simple explanation of intelligence merely on the grounds that what our brains - currently the best examples of intelligence - are doing *appears* to be very complicated*.

Contrariwise, and despite these optimistic examples, it is also logically possible that intelligence can only be explained by a large number of fundamentally distinct mechanisms. In the case of our brains, those many mechanisms may perhaps have evolved in response to many different selection pressures in our species' evolutionary history. If this point of view is correct, then intelligence involves considerable irreducible complexity, and no simple algorithm for intelligence is possible.

Which of these two points of view is correct?

To get insight into this question, let's ask a closely related question, which is whether there's a simple explanation of how human brains work. In particular, let's look at some ways of quantifying the complexity of the brain. Our first approach is the view of the brain from connectomics. This is all about the raw wiring: how many neurons there are in the brain, how many glial cells, and how many connections there are between the neurons. You've probably heard the numbers before - the brain contains on the order of 100 billion neurons, 100 billion glial cells, and 100 trillion connections between neurons. Those numbers are staggering. They're also intimidating. If we need to understand the details of all those connections (not to mention the neurons and glial cells) in order to understand how the brain works, then we're certainly not going to end up with a simple algorithm for intelligence.

There's a second, more optimistic point of view, the view of the brain from

Code repository

Michael Nielsen's project announcement mailing list

Deep Learning, book by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

cognitivemedium.com

By Michael Nielsen / Dec 2019

molecular biology. The idea is to ask how much genetic information is needed to describe the brain's architecture. To get a handle on this question, we'll start by considering the genetic differences between humans and chimpanzees. You've probably heard the sound bite that "human beings are 98 percent chimpanzee". This saying is sometimes varied - popular variations also give the number as 95 or 99 percent. The variations occur because the numbers were originally estimated by comparing samples of the human and chimp genomes, not the entire genomes. However, in 2007 the entire chimpanzee genome was sequenced (see also here), and we now know that human and chimp DNA differ at roughly 125 million DNA base pairs. That's out of a total of roughly 3 billion DNA base pairs in each genome. So it's not right to say human beings are 98 percent chimpanzee - we're more like 96 percent chimpanzee.

How much information is in that 125 million base pairs? Each base pair can be labelled by one of four possibilities - the "letters" of the genetic code, the bases adenine, cytosine, guanine, and thymine. So each base pair can be described using two bits of information - just enough information to specify one of the four labels. So 125 million base pairs is equivalent to 250 million bits of information. That's the genetic difference between humans and chimps!

Of course, that 250 million bits accounts for all the genetic differences between humans and chimps. We're only interested in the difference associated to the brain. Unfortunately, no-one knows what fraction of the total genetic difference is needed to explain the difference between the brains. But let's assume for the sake of argument that about half that 250 million bits accounts for the brain differences. That's a total of 125 million bits.

125 million bits is an impressively large number. Let's get a sense for how large it is by translating it into more human terms. In particular, how much would be an equivalent amount of English text? It turns out that the information content of English text is about 1 bit per letter. That sounds low - after all, the alphabet has 26 letters - but there is a tremendous amount of redundancy in English text. Of course, you might argue that our genomes are redundant, too, so two bits per base pair is an overestimate. But we'll ignore that, since at worst it means that we're overestimating our brain's genetic complexity. With these assumptions, we see that the genetic difference between our brains and chimp brains is equivalent to about 125 million letters, or about 25 million English words. That's about 30 times as

much as the King James Bible.

That's a lot of information. But it's not incomprehensibly large. It's on a human scale. Maybe no single human could ever understand all that's written in that code, but a group of people could perhaps understand it collectively, through appropriate specialization. And although it's a lot of information, it's minuscule when compared to the information required to describe the 100 billion neurons, 100 billion glial cells, and 100 trillion connections in our brains. Even if we use a simple, coarse description - say, 10 floating point numbers to characterize each connection - that would require about 70 quadrillion bits. That means the genetic description is a factor of about half a billion less complex than the full connectome for the human brain.

What we learn from this is that our genome cannot possibly contain a detailed description of all our neural connections. Rather, it must specify just the broad architecture and basic principles underlying the brain. But that architecture and those principles seem to be enough to guarantee that we humans will grow up to be intelligent. Of course, there are caveats - growing children need a healthy, stimulating environment and good nutrition to achieve their intellectual potential. But provided we grow up in a reasonable environment, a healthy human will have remarkable intelligence. In some sense, the information in our genes contains the essence of how we think. And furthermore, the principles contained in that genetic information seem likely to be within our ability to collectively grasp.

All the numbers above are very rough estimates. It's possible that 125 million bits is a tremendous overestimate, that there is some much more compact set of core principles underlying human thought. Maybe most of that 125 million bits is just fine-tuning of relatively minor details. Or maybe we were overly conservative in how we computed the numbers. Obviously, that'd be great if it were true! For our current purposes, the key point is this: the architecture of the brain is complicated, but it's not nearly as complicated as you might think based on the number of connections in the brain. The view of the brain from molecular biology suggests we humans ought to one day be able to understand the basic principles behind the brain's architecture.

In the last few paragraphs I've ignored the fact that 125 million bits merely quantifies the genetic *difference* between human and chimp brains. Not all our brain function is due to those 125 million bits. Chimps are remarkable

thinkers in their own right. Maybe the key to intelligence lies mostly in the mental abilities (and genetic information) that chimps and humans have in common. If this is correct, then human brains might be just a minor upgrade to chimpanzee brains, at least in terms of the complexity of the underlying principles. Despite the conventional human chauvinism about our unique capabilities, this isn't inconceivable: the chimpanzee and human genetic lines diverged just 5 million years ago, a blink in evolutionary timescales. However, in the absence of a more compelling argument, I'm sympathetic to the conventional human chauvinism: my guess is that the most interesting principles underlying human thought lie in that 125 million bits, not in the part of the genome we share with chimpanzees.

Adopting the view of the brain from molecular biology gave us a reduction of roughly nine orders of magnitude in the complexity of our description. While encouraging, it doesn't tell us whether or not a truly simple algorithm for intelligence is possible. Can we get any further reductions in complexity? And, more to the point, can we settle the question of whether a simple algorithm for intelligence is possible?

Unfortunately, there isn't yet any evidence strong enough to decisively settle this question. Let me describe some of the available evidence, with the caveat that this is a very brief and incomplete overview, meant to convey the flavour of some recent work, not to comprehensively survey what is known.

Among the evidence suggesting that there may be a simple algorithm for intelligence is an experiment reported in April 2000 in the journal *Nature*. A team of scientists led by Mriganka Sur "rewired" the brains of newborn ferrets. Usually, the signal from a ferret's eyes is transmitted to a part of the brain known as the visual cortex. But for these ferrets the scientists took the signal from the eyes and rerouted it so it instead went to the auditory cortex, i.e, the brain region that's usually used for hearing.

To understand what happened when they did this, we need to know a bit about the visual cortex. The visual cortex contains many orientation columns. These are little slabs of neurons, each of which responds to visual stimuli from some particular direction. You can think of the orientation columns as tiny directional sensors: when someone shines a bright light from some particular direction, a corresponding orientation column is activated. If the light is moved, a different orientation column is activated. One of the most important high-level structures in the visual

cortex is the orientation map, which charts how the orientation columns are laid out.

What the scientists found is that when the visual signal from the ferrets' eyes was rerouted to the auditory cortex, the auditory cortex changed. Orientation columns and an orientation map began to emerge in the auditory cortex. It was more disorderly than the orientation map usually found in the visual cortex, but unmistakably similar. Furthermore, the scientists did some simple tests of how the ferrets responded to visual stimuli, training them to respond differently when lights flashed from different directions. These tests suggested that the ferrets could still learn to "see", at least in a rudimentary fashion, using the auditory cortex.

This is an astonishing result. It suggests that there are common principles underlying how different parts of the brain learn to respond to sensory data. That commonality provides at least some support for the idea that there is a set of simple principles underlying intelligence. However, we shouldn't kid ourselves about how good the ferrets' vision was in these experiments. The behavioural tests tested only very gross aspects of vision. And, of course, we can't ask the ferrets if they've "learned to see". So the experiments don't prove that the rewired auditory cortex was giving the ferrets a high-fidelity visual experience. And so they provide only limited evidence in favour of the idea that common principles underlie how different parts of the brain learn.

What evidence is there against the idea of a simple algorithm for intelligence? Some evidence comes from the fields of evolutionary psychology and neuroanatomy. Since the 1960s evolutionary psychologists have discovered a wide range of *human universals*, complex behaviours common to all humans, across cultures and upbringing. These human universals include the incest taboo between mother and son, the use of music and dance, as well as much complex linguistic structure, such as the use of swear words (i.e., taboo words), pronouns, and even structures as basic as the verb. Complementing these results, a great deal of evidence from neuroanatomy shows that many human behaviours are controlled by particular localized areas of the brain, and those areas seem to be similar in all people. Taken together, these findings suggest that many very specialized behaviours are hardwired into particular parts of our brains.

Some people conclude from these results that separate explanations must be required for these many brain functions, and that as a consequence

there is an irreducible complexity to the brain's function, a complexity that makes a simple explanation for the brain's operation (and, perhaps, a simple algorithm for intelligence) impossible. For example, one well-known artificial intelligence researcher with this point of view is Marvin Minsky. In the 1970s and 1980s Minsky developed his "Society of Mind" theory, based on the idea that human intelligence is the result of a large society of individually simple (but very different) computational processes which Minsky calls agents. In his book describing the theory, Minsky sums up what he sees as the power of this point of view:

> What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle.

In a response* to reviews of his book, Minsky elaborated on the motivation for the Society of Mind, giving an argument similar to that stated above, based on neuroanatomy and evolutionary psychology:

> We now know that the brain itself is composed of hundreds of different regions and nuclei, each with significantly different architectural elements and arrangements, and that many of them are involved with demonstrably different aspects of our mental activities. This modern mass of knowledge shows that many phenomena traditionally described by commonsense terms like "intelligence" or "understanding" actually involve complex assemblies of machinery.

Minsky is, of course, not the only person to hold a point of view along these lines; I'm merely giving him as an example of a supporter of this line of argument. I find the argument interesting, but don't believe the evidence is compelling. While it's true that the brain is composed of a large number of different regions, with different functions, it does not therefore follow that a simple explanation for the brain's function is impossible. Perhaps those architectural differences arise out of common underlying principles, much as the motion of comets, the planets, the sun and the stars all arise from a single gravitational force. Neither Minsky nor anyone else has argued convincingly against such underlying principles.

My own prejudice is in favour of there being a simple algorithm for intelligence. And the main reason I like the idea, above and beyond the (inconclusive) arguments above, is that it's an optimistic idea. When it comes to research, an unjustified optimism is often more productive than a

seemingly better justified pessimism, for an optimist has the courage to set out and try new things. That's the path to discovery, even if what is discovered is perhaps not what was originally hoped. A pessimist may be more "correct" in some narrow sense, but will discover less than the optimist.

This point of view is in stark contrast to the way we usually judge ideas: by attempting to figure out whether they are right or wrong. That's a sensible strategy for dealing with the routine minutiae of day-to-day research. But it can be the wrong way of judging a big, bold idea, the sort of idea that defines an entire research program. Sometimes, we have only weak evidence about whether such an idea is correct or not. We can meekly refuse to follow the idea, instead spending all our time squinting at the available evidence, trying to discern what's true. Or we can accept that no-one yet knows, and instead work hard on developing the big, bold idea, in the understanding that while we have no guarantee of success, it is only thus that our understanding advances.

With all that said, in its *most* optimistic form, I don't believe we'll ever find a simple algorithm for intelligence. To be more concrete, I don't believe we'll ever find a really short Python (or C or Lisp, or whatever) program - let's say, anywhere up to a thousand lines of code - which implements artificial intelligence. Nor do I think we'll ever find a really easily-described neural network that can implement artificial intelligence. But I do believe it's worth acting as though we could find such a program or network. That's the path to insight, and by pursuing that path we may one day understand enough to write a longer program or build a more sophisticated network which does exhibit intelligence. And so it's worth acting as though an extremely simple algorithm for intelligence exists.

In the 1980s, the eminent mathematician and computer scientist Jack Schwartz was invited to a debate between artificial intelligence proponents and artificial intelligence skeptics. The debate became unruly, with the proponents making over-the-top claims about the amazing things just round the corner, and the skeptics doubling down on their pessimism, claiming artificial intelligence was outright impossible. Schwartz was an outsider to the debate, and remained silent as the discussion heated up. During a lull, he was asked to speak up and state his thoughts on the issues under discussion. He said: "Well, some of these developments may lie one hundred Nobel prizes away" (ref, page 22). It seems to me a perfect response. The key to artificial intelligence is simple, powerful ideas, and we can and should search optimistically for those ideas. But we're going to

need many such ideas, and we've still got a long way to go!

In academic work, please cite this book as: Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015

Last update: Thu Dec 26 15:26:33 2019