A conversation with Joscha Bach, 9/21/2015

Interview by Katja Grace and John Salvatier of AI Impacts
Summarized by Connor Flexman

Summary

- Before we can implement human-level artificial intelligence (HLAI), we need to understand both mental representations and the overall architecture of a mind
- There are around 12-200 regularities like backpropagation that we need to understand, based on known unknowns and genome complexity
- We are more than reinforcement learning on computronium: our primate heritage provides most interesting facets of mind and motivation
- AI funding is now permanently colossal, which should update our predictions
- AI practitioners learn the constraints on which elements of science fiction are plausible, but constant practice can lead to erosion of long-term perspective
- Experience in real AI development can lead to both over- and underestimates of the difficulty of new AI projects in non-obvious ways

I When will HLAI come?

We don't have the specification for an AI architecture yet, so any estimate of an HLAI due date may be off by orders of magnitude. There is a remote possibility there is a silver bullet just around the corner.

However, minds can't be *that* difficult because we can fit the genome sequence on a CD-ROM, and the genome contains the instructions for the whole brain. The Kolmogorov complexity is simply not that high.

Then it seems like there are only up to a few hundred regularities that we need to understand, each perhaps of the difficulty of backpropagation, such that it is easy to understand in hindsight. We know maybe 20 of these so far. From this, it seems plausible to be done in fifty or a few hundred years, perhaps most likely a hundred.

How do you determine the number of regularities left?

The simplest way is to ask a geneticist about the number of functional regularities encoded in the fraction of the genome that codes for the brain. One can also keep up with the number of currently known problems and extrapolate. Something must govern the limbic system, and all the nuclei, and re-routing, and cortical functionality, so there are probably more than a dozen principles.

What are the major unsolved problems?

By no means an exhaustive list, but to get to HLAI we likely must solve the problems of

- defining a universal network that can account for any kind of perceptual content
- the relationship between content and percepts
- the relationship between content and natural language
- a mental stage to present counterfactuals and arbitrary simulations and expectation (imagination)
- communication to oneself
- communication of ideas to others
- language formation

- learning in the general case
- the motivational system (not very difficult)
- cognitive and physiological and social goals
- memory allocation and management (incl. forgetting)
- an attention regulation system distributing over cognitive resources
- decision-making with resource constraint
- top-down / bottom-up processing over different sensory modalities
- reflection
- activation modulation
- near-universal problem solvers (some problem classes are very hard for us; there may be some impossible ones we don't know about)

II What are some large knowledge gaps on the road to HLAI?

Overall Architecture

Many of the regularities discovered should be architectural principles: which parts of the mind do what, and in what proportion. This includes computational constraints on different parts of the system. Much effort has gone to understanding the architecture of the visual cortex, but unfortunately less so to the other parts of the brain. The blueprint that shows how an AI should fit together will be vital.

To define how something works, you need to specify both what all of the parts do and how all of the parts fit together. For example, once we know how muscles work, we can replace them by something that plays the same role, like motors. However, you still need an overall architecture dictating where the muscles attach if the system is going to function properly.

In the brain, there is a similar architecture we need to determine. We could replace cortical columns with something that is computationally equivalent (like Minsky's "schemas"), but we still would need to determine the overall architecture with which they are connected. Our brain has a lot of functional differentiation. It isn't some kind of neural sponge that self-organizes if you put a spine on one end and eyes on the other end and expose it to the right kind of data flow. All classes of organisms have feet on one end and eyes on the other, but they aren't smart in the same ways.

In fact, we can partially study this by observation. There are 3 strong brain designs so far: bird, octopus, and human. Our brains are very expensive, requiring 20% of our energy, so you'd think there would be a very strong evolutionary pressure to optimize it. However, birds are extremely smart with a tiny fraction of our brain size and the same overall architecture. The jury is out on whether the difference is qualitative or quantitative, since humans are the only ones to develop language but both others are very good at planning and problem solving. Octopus are especially interesting, as they have very different brain development history and thus a different mesoscopic architecture. However, much of it has functional equivalence with humans. It would be good to study this in more detail.

Mental representations

Most of the non-architectural regularities discovered will likely bear on mental representations. The systems built with these new insights will be qualitatively closer to HLAI.

One milestone will be the advent of imagination and microsimulation so that possible worlds can be constructed after capturing the main constraints. Once mental representations are understood, a system might be built that can be fed a photograph of an Italian village and subsequently construct a 3D model with believable housing interiors. Perhaps the system is told that the village center has been replaced it with a modern office building, and it will output what that looks like.

Another milestone will be arbitrary language learning: capturing the difficulties you have in translating mental representations of directed quasi-hypergraphs into discrete strings of symbols with limited resources. If you can do this in the general case, you can probably join all natural languages.

Primate heritage in intelligence, values, and reflection

Schmidhuber advances the notion that reinforcement learning (RL) fully explains the complexity of our brains, but I disagree. To build an HLAI, it seems very difficult to just brute force it in the RL regime. Our intelligence isn't like an arbitrary rational agent, but a primate that got rational. This makes our minds interesting and gives them structure.

Our values don't look like an RL agent either. Primate goals are what drive all of our interesting behavior, like communicating, socializing, problem solving, trying to get out of bed, and searching for the right level of environmental complexity. These values make rationality useful in particular ways, like creating compelling social arguments. If you removed all constraints and modeled us only as computronium with a single goal, we'd probably look more like a wildly efficient virus.

RL also seems to insufficiently explain reflection. You can expand the notion of RL toward imagination, mental stimulation, and all the things that occur during meditation and deliberation, but that doesn't seem like a productive framework when aiming for all these associated faculties. An organism that reflects almost certainly performs better than one that doesn't, but it's not clear that reflection would arise directly from training an RL system, especially current ones. Reflection would seem to require a separate mechanism, either built-in or found by a giant evolutionary search.

III How can we improve predictions about AI?

Appreciate the impact of funding

Perhaps the biggest factor that should update our predictions is the quickly increasing state of AI funding. The current generation of IT CEOs believe in AI, and the funding won't stop. The sheer level of it is a major impetus that people didn't expect in 2010, much less the '70s.

Even so, with diminishing returns the number of results may not be accelerating. Many of the deep learning ideas we are testing now were already conceived in the '80s and '90s, and we simply didn't have enough hardware capacity to use productively. It's not clear how long until the next generation of ideas.

Acknowledge constraints

The first AGI conference had many science fiction fanatics whose view of the impacts of AGI was inversely proportional to their involvement with practically implementing AI. This might be because such excitement is unwarranted, or because involvement with day to day problems erodes practitioners' long term perspective. But as a practitioner, experience leads to an understanding of constraints on the space of actual possibility.

Accept nuance in factors altering estimated timelines

Outsiders frequently overestimate or underestimate the difficulty of various problems in AI because of the incomplete information they receive. For example, those working in other subfields of AI will often hear of the difficulties arising in self-driving car design when a GPS fails or a road is cracked. However, they won't see as many of the solutions, which are black-boxed by the engineers who solve them. This can lead to overestimates of the difficulties involved. On the other hand, the public may see constant steady progress and underestimate future difficulties.