A Conversation with Tom Griffiths

Participants:

- Professor Tom Griffiths - Director of the Computational Cognitive Science Lab and the Institute of Cognitive and Brain Sciences at the University of California, Berkeley.
- Finan Adamson - AI Impacts

Note: These notes were compiled by AI impacts and give an overview of the major points made by Professor Tom Griffiths.

Summary:
Professor Tom Griffiths answered questions about the intersection between cognitive science and AI. Topics include how studying human brains has helped with the development of AI and how it might help in the future.

How has cognitive science helped with the development of AI in the past?

AI and cognitive science were actually siblings, born at around the same time with the same parents. Arguably the first AI system, the Logic Theorist, was developed by Herb Simon and Allen Newell and was a result of thinking about the cognitive processes that human mathematicians use when developing proofs. Simon and Newell presented that work at a meeting at MIT in 1956 that many regard as the birth of cognitive science - it was a powerful demonstration of how thinking in computational terms could make theories of cognition precise enough that they could be tested rigorously. But it was also a demonstration of how trying to understand the ways that people solve complex problems can inspire the development of AI systems.

How is cognitive science helping with the development of AI presently?

When I think about this relationship, I imagine a positive feedback loop where cognitive science helps support AI and AI helps support cognitive science. Human beings remain the best examples of systems that can solve many of the problems that we want our AI systems to solve. As a consequence, insights that we get from studying human cognition can inform strategies that we take in developing AI systems. At the same time, progress in AI gives us new tools that we can use to formalize aspects of human cognition that we previously didn't understand. As a consequence, we can rigorously study a wider range of questions about the mind.

How can cognitive science help with the development of AI in the future?

Deep Learning Systems

Deep learning systems are mastering a variety of basic perceptual and learning tasks, and the challenges that these systems now face look a lot like the first important stages of cognitive development in human children: identifying objects, formulating goals, and generating high-level conceptual representations. I think understanding how children do these things is potentially very relevant to making progress.

Efficient Strategies

One of the things that people have to be good at, given the limited computational capacity of our minds, is developing efficient strategies for solving problems given limited resources. That's exactly the kind of thing that AI systems need to be able to do to operate in the real world.

What are the challenges to progress in studying brains as they relate to AI?

Birds and Planes

One important thing to keep in mind is that there are different levels at which we might see a correspondence between human minds/brains and AI systems. Critics of the idea that AI researchers can learn something from human cognition sometimes point out that the way jet airplanes work has little relationship to how birds fly, and in fact trying to mimic birds held back the development of planes. However, this analogy misses the fact that there is something important that both jets and birds share: they both have to grapple with aerodynamics. Ultimately, we can see them both as solutions to the same underlying physical problem, constrained by the same mathematical principles.It isn’t clear which aspects of human brains have the best insights that could cross over to AI. Examples of places to look include the power of neurons as computational units, the efficiency of particular cognitive strategies, or the structure of the computational problem that is being solved. This last possibility — looking at abstract computational problems and their ideal solutions — is the place where I think we’re likely to find the equivalent of aerodynamics for intelligent systems.

What blind spots does the field of AI have that could be addressed by studying cognitive science?

I don't think they're blind spots, they are problems that everybody is aware are hard - things like forming high-level actions for reinforcement learning, formulating goals, reasoning about the intentions of others, developing high-level conceptual representations, learning language from linguistic input alone, learning from very small amounts of data, discovering causal relationships through observation and experimentation, forming effective cognitive strategies, and managing your cognitive resources are all cases where we can potentially learn a lot from studying human cognition.

How does cognitive science relate to AI value alignment?

Theory of Mind
Inferring the preferences or goals of another person from their behavior - something that human children begin to do in infancy and gradually develop in greater sophistication over the first few years of life. This is part of a broader piece of cognitive machinery that developmental psychologists have studied extensively.

What risks might be mitigated by greater collaboration between those who study human brains and those building AI?

We're already surrounded by autonomous agents that have the capacity to destroy all human life, but most of the time operate completely safely. Those autonomous agents are of course human beings. So that raises an interesting question: how is it that we're able to create human-compatible humans? Answering this question might give us some insights that are relevant to building human-compatible AI systems. It's certainly not going to give us all the answers - many of the issues in AI safety arise because of concerns about super-human intelligence and a capacity for self-modification that goes beyond the human norm - but I think it's an interesting avenue to pursue.