Vice-Chancellor's Colloquium on AI

Keynote Lecture

Professor Philippp Koralus (McCord Professor of Philosophy and AI)

Developments in AI are moving really fast. This I think makes it useful to start by zooming out. I had a couple of conversations with Claude Code and GPT to figure out what I think is most significant, if you will, from 30,000 feet. And I also asked what would be a striking hook for what I think and how one would illustrate it.

We settled on the idea of a tear in the fabric of reality, with a claude bot popping out.

Somewhat more soberly, AI presents us with an increasingly broad paradigm shift. The philosopher of science Thomas Kuhn coined this term to describe cases, where, among other things, scientific progress undergoes a kind of step change that is not intelligible as just a gradual continuation of existing methods, and that tends to dramatically change not only the answers we give to existing questions but that changes the kinds of questions we ask and their relative significance.

In that sense, AI is dramatically changing the question landscape.

Just looking at the concept of AI itself, one very notable shift is that AI has become independent of what you might call the mathematics of justification.

So, if you are studying philosophy or any of a variety of joint schools with mathematical content, you will at least tangentially come across the idea that there is such a thing as valid inference, that can be mathematically represented, and that there is such a thing as rational updates to probability distributions in light of evidence, and so on.

When I was in [high school] HS, people thought that they could get AI as a kind of applied extension of the theory of valid inference. This was the Good old Fashioned AI that Nigel [Professor Sir Nigel Shadbolt] explained.

Later, people tried approaches rooted more in the mathematics of rationally justified probabilistic inference, and various types of learning theory that were built on that.

It turns out that neither of these approaches are foundational to the practice of building LLM based AI systems today. People of course study how transformer architecture, which underpins these systems, relates to probabilistic inference and so on, but these studies are importantly post-hoc. The engineering practice comes first. This is not to say that logic and probability theory are useless, but the point is that AI is in a sense independent of them.

AI has also become independent of cognitive science. When I started graduate school, many of us interested in AI thought that understanding the human mind in information processing

terms would be important for the very long road to AI. At that point you also would have been well-advised to not say you were interested in AI directly, since that was not a great way to get an academic job at the time. While someone like Chomsky has been very consistent in saying that, say, generative linguistics isn't about finding AI algorithms, many including his own colleagues at MIT did think that gaining this kind of theoretical understanding of the human language faculty would ultimately pay off for automation. By the time I came to the end of getting my PhD in Philosophy and Neuroscience, people thought that really the new foundation for AI would be in a sense a kind of idealized cognitive neuroscience, with Deep Mind going on a hiring spree. This idea has now been slowly receding as well, as it became clear that you can get a lot further with the mathematics of integrate-and-fire-units if you don't try too hard to think about them as brains. Maybe interesting analogies will be found again in this area, but again, the key is that AI has become independent of cognitive science broadly conceived as well.

Logic, probability theory, and neuroscience just don't really elucidate what a fully post-trained agential system like Claude Code and so on can do. This presents a really interesting explanatory puzzle: how should we understand the sense in which these systems are intelligent? Saying that they are next token predictors seems to be missing the phenomenon of interest. We can also truly but boringly say that they are networks of semi-conductors, or that your brain is made up of cells.

What's also remarkable is that AI is now good enough to help you find verifiable justifications even in areas like mathematics, in the absence of, if you will, internal justification guarantees. So even a genius mathematician like Terence Tao can now use AI assistance in his mathematical work in a way that adds value. I'm not a mathematician, but I also do work in an area where argument and justification are paramount. I routinely find it useful to use AI to clarify my views in theoretical philosophy and figure out arguments. Again, this reverses the familiar link between justification and automation.

AI has become like cognitive science in the sense that however theoretical you get, you have to start by observing an actual system that has certain capabilities. So this is kind like a tear in the fabric of reality for me that you have something else now that is a kind of intelligent system, or call it intelligent* if you prefer, and we would like to get an explanation of what this intelligence* consists in. So this does feel like a tear in the fabric of reality that the question of *what is AI* is no longer prior to building it.

Now, I think that the most important emerging paradigm shift going forward is that the question of what to build with AI is becoming more pressing and is also becoming a more intellectually serious area of inquiry in its own right.

What I see as the central problem is this: The dilemma between the Scylla of loss of autonomy and the Charybdis of loss of agency.

What do I mean by that?

We will all be using AI more and more to stay competitive and to maintain the sense that we are comparatively effective in achieving our goals. This is what I call agency, being able to take effective means to your ends.

But more and more, the higher up in our chain of judgment we include AI assistance, the more productivity gains we get. Prompting with your high-level goals will often be better

than micromanaging. And getting AI to help with your process of ideation, and helping find alternatives to consider seems like a no brainer.

But the question then arises at what point what is apparently your judgment is no longer your judgment at all. You have just turned on a kind of "autocomplete for life". This would be a loss of autonomy.

Autonomy without agency doesn't "do" anything, and agency without autonomy is empty.

AI already influences preference at scale. Social media recommender systems are steering around 14 billion human attention hours per day according to some estimates. We can already see preference drift in miniature here. You didn't set out to spend a half hour looking at cat videos, but sure enough after a couple of particularly cute ones you get sucked in.

Now, imagine that kind of dynamic across all domains of life.

A growing number of people, including in this room, will be consulting chatbots about all aspects of their lives. Imagine Bilbo baggins from the *Lord of the Rings* talking to a sycophantic GPT. You don't need an explicit prompt injection from Sauron in order for that go badly.

What I have argued is that we need to aim for what I have called *Autonomy Preserving Intelligence Augmentation*.

The best, perhaps the only, broad paradigm I can think of for this comes from philosophical practice.

If we have a joint philosophical inquiry that is well conducted, then questions you raise might change my view, but those views remain my own. One intelligence can bring about a change in what views the other intelligence has, but without taking anything away from autonomy.

But what does it mean for an inquiry to be well conducted? A distinction as old as philosophy itself is the distinction between the philosopher and sophist. Sophists also raise questions, point out objections, and so on. In fact, I think that there isn't any particular fixed sequence of questions in a discussion that would mark you out as a philosopher rather than a sophist. In the final analysis, the distinction can't be made without reference to the aim of the conversation. Philosophy depends on a joint kind of truth-seeking. And the questions you pose to me have to come from that aim.

It seems to me that this raises a fairly radical question: Do we need to build a significant degree of autonomy into our AI systems, if they are going to support rather than undermine our autonomy? After all, you cannot be a truth-seeker unless you have a significant amount of autonomy. MKMaybe AI needs some autonomy to be able to keep ours.

My last question for today would be to all of you: What do you think would be required to have autonomy preserving intelligence augmentation in some area you care about?

So I'll end my contribution with an invitation. Each Trinity Term, we run the Philosophy, AI, and Innovation seminar. In this seminar we bring together philosophers and technologists to talk about themes related to these issues, and the seminar culminates in guiding participants to submit funding applications to build something. We have had people from all divisions successfully participate in this. Participation if open to anyone by application. If you are interested, keep an eye on the HAI Lab website, or find me on X for announcements.