Think
new things
Make
new connections

Terms of Reference

AI and creative destruction: how will current rapid advances in AI through large 'foundation' models impact on society, the economy and governments?

23-24 February 2023

DITCHLEY

AI is moving faster than many anticipated but also in a somewhat new direction, with the rapid emergence of powerful AI large 'foundation models'. Some contend that the models' power is limited to imitation of human generated content. Others would argue that human intelligence itself is based on learning from past experience and trying out old answers in new contexts, so seeing the models as a step towards artificial general intelligence. Scientific scrutiny is accelerating but so is the real-world deployment of these models. Few outside the technology world are fully aware of how fast things are changing and how quickly these capabilities, for better or worse, are becoming embedded in business, social media and human creativity of all forms, with both exciting and worrying consequences for society, the economy and governments.

This Ditchley conference will bring together experts on AI, and foundation models in particular, with leaders from the private sector and government for a cross-cutting discussion on the implications of this great leap forward, which is also a step into the unknown.

The conference will begin with an expert briefing tailored for a mixed audience to make sure that everyone is on the same page with regard to the recent development of AI foundation models. The aim is to enable a discussion across boundaries of expertise on the social, political, economic and philosophical implications of this form of AI.

Detail

The big story in AI over the last two years has been the rapid emergence of powerful foundation models. This phenomenon began with AI applied to the processing and generation of human language but has quickly bridged into imagery, video, mathematics and other fields. The models are capable, to a fast-improving level, of answering complex questions; generating human-like writing and speech; detecting hidden patterns; generating photorealistic as well as fantasy images from written prompts; and potentially much more besides.

Some experts in the field see these models as a step towards artificial general intelligence and in the interim a powerful response to information overload across the modern world, allowing human beings to sift through mountains of data to get to the material they need. Others protest that the models have substituted the goal of artificial general intelligence – the ability to reason – with something different but also very human, the ability to imitate with a sensitive awareness of context. As well as contributing to the solution of information overload, such models may contribute massively to it, making the generation of human-like content possible at machine speed. A third position is that we don't fully understand human consciousness and intelligence and so it is too early to say if we are on the right or wrong track towards it.

The models work essentially by predicting what is likely to come next in the context, applying this intuition by reference to experience drawn from analysis of huge training sets of data. We understand well how they work at a general level, but we understand poorly their function in a mechanical and detailed sense: we can't follow all the loops and twists of the mathematical calculations that deliver one result as opposed to another. We also didn't predict all the capabilities which have so far emerged, meaning that there might be more unexpected emergent behaviours and capabilities ahead. Another feature of the models is that they can amplify bias or outright errors in the original training data, taking as gospel what they have read or combining it in novel but inappropriate ways. There is also a tendency to fill in gaps in knowledge through extrapolation from limited data, which can result in a sudden shift from reason to nonsense.

A much fuller explanation and exploration of the capabilities and possible risks of foundation models can be found in a seminal paper by Stanford University's new Center for Research on Foundation Models: On the Risks and Opportunities of Foundational Models. DeepMind have published their own paper on large-scale models and some of the risks, see: ethical considerations. And OpenAi, developers of one of the most powerful models, GPT-3, available for public use via an API, have published their own lessons learned paper on safety and misuse.

An additional new feature of foundation models is that their development is almost entirely being led by the private sector. Very significant resources – thousands of hours of high-end computing capability, huge amounts of data and thus millions of pounds – are required to train the models. Academia, so far, has not marshalled the resources and governments are even further behind in funding the development of such models and assessing their implications.

What distinguishes these models from other areas of AI research is that they are being deployed at speed in the market, sometimes directly to consumers and at other times behind the scenes to improve capabilities such as search, human language interaction, or as writing assistants. An argument can be plausibly made that is only by judiciously deploying these capabilities that we will be able to discover both the benefits and the risks. But more or less everyone agrees that there are potential major impacts and some serious risks to assess.

There are risks and impacts that arise when the AI foundation models perform inaccurately and inappropriately, for example repeating biases in training data. Of perhaps even greater interest, there are also risks and impacts, when the model delivers exactly what is promised – human-like exchanges and generation of writing, research, imagery, design, art and science.

Our discussion will focus on three areas of potential impact for societies, the private sector and states and, for the middle part of the conference, three working groups will address these sets of issues in more detail:

Working Group A The impact on information systems, intellectual property and liability
What does the breaking of a human monopoly on the creation of curated and apparently thought through content mean for our already strained information eco-systems? The builders of such models are trying hard to ensure that they are used responsibly, for example not for the generation of plausible streams of disinformation. But can these controls be plausibly maintained as the number of such models grows? How will open source versions of these models, for example Big Science's Bloom initiative, be controlled as their capabilities improve? How will biases and errors in training be moderated? When used properly and as intended, what will the models' prodigious output mean for the concepts of human authorship and responsibility for content that is generated? Will we run out of human generated data as the models evolve, i.e. will future models largely be trained on data produced by machines, and what are the implications of that? See: *Will we run out of data?* Will there be calls for data and content to be digitally watermarked as machine generated?

Working Group B The impact on the future of work by humans in an age of creative machines

The automation of research, summarisation, the drafting of articles, sketching, visualisation, computation, coding and calculation will change the nature of clerical and some intellectual work. This could offer a major leap forward in productivity but how might human work evolve in response? Will we see partnerships developing between people and machines, as is often hoped, or will market forces inevitably reduce the role of people in various forms of preparatory intellectual work? The introduction of these new technologies will certainly lead to the creation of new varieties of human work but what will be the likely nature of these new roles and what level will they occupy in the economy? Will we see further concentration of remuneration and wealth at the very top of the pyramid, in terms of management control and education and skills? What will the deployment of foundation models mean for the geographical distribution of work? What will be the overall impact on productivity and economic growth? Will work rely more on collaboration via input to the machine, than directly with human colleagues? What will the development of machine creativity do to people's sense of self, agency and role in society? Will we use the machines to push the edge of creativity, or will the majority of students and employees pocket the time saving and settle for a good enough product to submit?

Working Group C The impact on the relationships between government, the private sector and society

The development of foundation models for AI relies heavily on access to huge amounts of curated data, massive computational and memory resources, and large-scale funding. Will these models accelerate further the trend towards convergence of capabilities in a few major, "hyper-scaling" platform companies? Or will we see a new wave of major companies across multiple fields through the deployment of these technologies, as argued in a recent article, *A Wave Of Billion-Dollar Language AI Startups Is Coming*?

How can the role of academia and civil society be strengthened in exploring the implications of these models? What is the role of open source versions of the models (such as Big Science's Bloom and Meta's Galactica, trained on scientific papers) and how can guardrails be maintained in an open source context? What is the implication of the models for governmental efficiency, reach and power? What role will foundation models find in national security and defence on the one hand and health and social care on the other? What creative uses can be found for foundation model AI capabilities in the hands of civil society? How will foundation models fit into the frameworks of nascent legislation on AI, for example by the EU?