Think
new things
Make
new connections

Conference Summary

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

EXECUTIVE SUMMARY

Context and why this was important

February 2023 was a moment when suddenly, thanks to the viral success of Chat GPT, the rapid progress in AI through large language models burst into the consciousness of wider society. After years of complaints that AI wasn't working, suddenly it seemed it was, with the arrival of the kind of AI that we've been promised in films and science fiction, with powerful and, superficially, perceptive capabilities.

Open AI, with support from Microsoft, had stolen a march on the other major tech companies and that success was triggering some genuine alarm about possible consequences, mixed with some regretful jealousy that others had not moved earlier. The success of ChatGPT accelerated a major pivot towards generative AI, with Google and Meta all releasing their versions within weeks and venture capital flowing to accelerate an effort already moving forward exponentially. Corporate dominance was at stake and in such a race there was a concern that all other considerations, including potential risks for society, could become secondary. Governments were clearly utterly unprepared for the speed of change which might be unleashed, with legislation in the pipeline geared towards the data use and privacy implications of current internet uses, not radical new challenges.

People

Many of the leading start-ups and major technology companies involved in the rapid development of generative AI were represented in the discussion. They were joined by senior scientists, commentators, venture capital investors and government.

Summary of the discussion

This was a group deeply vested in generative AI and with great collective knowledge. There was clear enthusiasm for **the potential of such AI to deliver greater productivity, to free people from drudgery, to strengthen national capabilities and to deliver returns on investment**. But it was striking that where the discussion returned time and time again was to the likely disruption that rapid change would bring and other potential risks, especially for the technology to unleash unintended real-world consequences, and perhaps to spin out of control. There was a sense of being engaged in the making of a fast-emerging new world and a personal responsibility to speak up on the negative possibilities. But at the same time the excitement and human temptation to open the next door in Bluebeard's castle was palpable.

There are threats to democracy from unintended consequences which include **potentially a further exacerbation of inequality and perhaps also a further undermining of agreed facts and common truths.** This was described as the most important technological change of our lifetimes with massive investment of capital and talent and yet government (UK, Western) capacity to understand and respond is currently very limited.

The conference began with a briefing session to share understanding of the way these large language models operate, with demonstrations of a powerful chat interface able to answer general questions, calculate and summarise in different styles and the deployment of AI speech capabilities via telephone for care calls to elderly citizens in Korea. The Korean AI agent had been given access to records of its previous calls which meant exchanges that were both personalised and, to many listening, somewhat uncanny compared to the robotic exchanges we are used to with smart speakers: "Are you still suffering with your stomach pain, Mrs Kwan?"

Large language model code was described as grown rather than written. We do not know, let alone have the ability to track, its full complexity. As a result, alignment – the AI's action matching what the human initiator intended – was a deep and fundamental problem that would not be solved easily. On the other hand, there was a clear continuing exponential trajectory over the next two years in terms of general performance. If the models were average students in most exams now, we could expect them to be outstanding very shortly and, in some fields, quickly to be beyond human performance, albeit with that alignment problem, meaning that veering off the tracks from time to time would remain inevitable. Not only would the models continue to get things wrong unexpectedly but they tended to be very confident in defending such errors.

Large Language Models (LLMs) have developed some unexpected emergent functions, with capabilities arising to developers' surprise. More significant, perhaps, is the potential to create code to add to other systems. This means that LLMs could be linked to legacy software such as API databases and then in time to forms of embodiment – i.e. robotics allowing actions in the physical world. It was expected that LLMs could become an interface layer between humans and legacy software, allowing chains of automation. Alongside novel functionality in some directions were surprising limitations in others and there was debate about whether this technology can produce new knowledge or just recombine existing knowledge. Or, does this form of AI challenge our understanding of what knowledge is? For example, it can demonstrate advanced computation without any specific knowledge of maths. What then is maths? Is it just syntax, rather than semantics?

Generative AI has emerged at time of contested geopolitics and this means heightened attention to the links to economic security and the risks of over-reliance on foreign technologies and supply chains. International competition between states is compounded by competition between corporations. It is hard to know how the ownership of such a powerful technology will play out in the years ahead in the context of strategic competition between the West and China.

Is it possible to apply existing regulation or to regulate according to particular use cases, i.e. with attention to the applications rather than the technology itself? Would it be better to make the owners/companies/governments apply ethical principles than to attempt to inject ethical components into a technology that at its core is neither inherently bad nor good? Or will this technology be let loose and ramp up existing algorithmic trends that drive social fragmentation?

At the same time, institutions and frameworks capable of enforcing effective safety mechanisms and rules are either blocked, weak, uninformed or just not yet conceived. Increasingly (and across multiple domains), effective global regulators are absent. It was in this context that some made the argument that generative AI could not currently be effectively regulated and that therefore a slowdown in its development and roll out was the only responsible step. At the same time, though, no one really expected that to happen. As a result of this ground-breaking technology being developed entirely in the private sector and with an eye to markets, we are about to hit a major acceleration, with change coming much faster than we are accustomed to. How will human societies manage this and are there new kinds of mechanisms to help us alongside or beyond regulation?

Questions about how humans maintain control and the capability to defend human interests were seen as pressing. Which humans stand to be the winners and who will lose out as this technology and the corporate competition that frames its delivery unfolds? Can current regulatory systems be stretched to accommodate this novel and broadly applicable technology, or is a different and more radical approach required? One radical approach was championed by some in the discussion: the creation of personalised open-source generative AI agents so that each country, company and individual will have some leverage, and their productivity and influence boosted. If the cat was out of the bag, then democratisation and proliferation of the capability was put forward as the only answer.

Recommendations and ideas (not consensus but arising in the discussion)

Accept that this is new and different. There was thought to be a genuinely new level of innovation at play, characterised by flexibility of application. Models can be adapted from one context to another and new pathways, new parts of models, potentially meaning new emergent capabilities could be unlocked as they interact with the world. We have invented a 'machine' but we don't know yet all the things that it can do, or might do. What should be the early warning tests that we are losing control over a process? How should we measure and test as we go, in order to take action and develop collective understanding about safety and what works?

Create government-controlled labs for the most edgy work on AI. We were urged by some of those at the conference not to dismiss the risks of the accidental creation of artificial general intelligence (AGI), or a step towards it which might run out of control, for example by causing widespread damage to the Internet. Some argued that the time to AGI might be as little as zero to three years. Other scientists were more sanguine, and this was repeatedly debated.

Start with an assessment of what the impacts are likely to be in the knowable short term. Which job roles will be hardest hit? What forms of discourse and online exchange will be most affected? How can the models help separate fact from fiction and improve search and productivity? Urgent work is needed.

Map what might happen to different roles in the workforce. It is likely that a percentage of each job could be automated by a generative AI. Can we map the roles where that percentage will be high in the short term and develop plans to address re-training and transition?

Develop the right talent. What kind of talent and what kind of education is now required for us to both further develop and also manage or work alongside generative AI? How can generative AI transform education and training to enable us to deliver that talent more quickly?

Define the potential guardrails. Are there lessons to be drawn on from previous efforts at democratic oversight over technology, for example biosafety, genetic engineering or nuclear weapons? Can we learn more from open-source precedents along the lines of Linux and Wikipedia, or from closed models such as CERN, *Los Alamos* National Laboratory or ITER (international nuclear fusion research)? The US Food and Drug Administration was suggested as a possible model, with some supportive, others sceptical about its promise.

Make disclosure an obligation. Can initial regulation be passed to make disclosure on safety an obligation? There was a tension between broad regulation measures and specific use cases. Can solutions be found for specific use cases building on existing practice in any given field, or is broader and more fundamental regulation of the technology needed?

Educate policymakers. The state needs rapidly to acquire a better understanding of the power of these models. How can we quickly educate policymakers and politicians?

Inputs to existing processes and draft legislation. Current draft legislation around the world, for example the draft EU AI Act, may be unfit for generative AI's general capabilities and relevance. What advice can be put forward for policy on regulation and for investment in research and development?

Define what it means to develop sovereign capabilities. Should governments have a role in trying to assure national access to the fundamental hardware underpinning generative AI – large clusters of tens of thousands of high-level GPUs? What does this mean for industrial policy? What government

funded research is needed? Do governments need to train their own generative agents, creating 'sovereign' large language models that represent national values, not those of others?

How will benchmarking and evaluation be achieved? *Being grown rather than written*, generative AI code cannot be scrutinised, debugged and audited in the same sense as normal code, even that drawn from libraries. Emergent behaviours cannot be known in advance. Is it possible to develop systems to provide guardrails on a step-by-step basis? Can we incentivise such checks by introducing liability for developers (companies) when they deploy these models?

What are the human rights that apply and do we need new rights to protect us from new threats? Do we have a human right to a human decision? What rights do we need to protect our identity and reputation when our online presence can be convincingly created now in words and images and soon in sound and video? Who must lead in protecting these rights?

Is the idea of an 'ethical or responsible AI' a distracting fantasy? Would it be better to attempt a more structural shift to align technology with the interests of humanity? In other words, to focus on ensuring ethical companies rather than ethical generative models.

How can the vulnerable, and society and individuals more broadly, be protected? What happens to us (positively or negatively) when the output of generative AI models becomes our information environment? And in due course when their output becomes largely their own training data, rather than content solely created by human beings? Could manipulation of users reach a new level? For example, communication to us in our own 'voice' and our own patterns of 'language' could be harder to resist. And who will own the newly created or 'co-created' outputs made between humans and commercially provided generative AI? How will users understand these new content 'forms'? What should be the special protections for children and others who are most vulnerable to manipulation?

Will generative AI accelerate the existing trend towards the segmentation and tribalisation of society by technology? Will the delivery of highly personalised content in our own voice further split us into different interest groups? Could the models address problems of disinformation and misinformation, for example through better search, summarisation and fact checking, or make them much worse, for example through streams of convincing but fake news and images generated at machine speed? What lessons can be learned from the failures and mistakes of existing efforts for platform moderation? Providers and regulators will have to understand both the wider ecosystem and the scope for personalisation, even hyper personalisation and the consequences of new levels of fragmentation.

Several of the themes and questions arising from this conference will inform Ditchley's programme for 2023-25. These include:

- The implications of uses of AI in defence technologies.
- How to foster resilient innovation in the context of geopolitical competition.
- How best to develop UK capabilities and compute resource.
- How best to bring together expertise to support assessment of key risks and opportunities of generative AI and foundation models.
- What will the impact of generative AI models be on our news and information systems?
- The impact on 'work', jobs and the identification of future skills.
- Economic in/security as a risk and policy goal.
- Data uses, constraints and opportunities within systems in the Global South.
- Ditchley will experiment with the functionality of generative models in the delivery of the Ditchley method and mission.

Further detail

Human decisions, public infrastructure and the power of open source

The arrival of a technology able to generate plausible content, whether text, images, video or speech, raised philosophical questions about challenges to human decision-making and what the input of this kind of technologically generated content might mean for human judgement, responsibility and trust. Could new uses change our understanding of human rights and what does the AI-generated content do to rights to privacy? How private is our interaction and what might AI-generated content say about us that is untrue? What recourse do we have and what responsibilities do tech companies have to enforce human rights governance? What are the genuinely novel areas for which new kinds of regulation will be necessary?

There was debate about whether this technology could or should become a public good and whether open source could help to define, deliver and certify digital public goods and the public infrastructure necessary to ensure equal access. Would a lean customisable open-source model with reduced costs allow greater access and open the technology, for example to the global South, so creating opportunities to leapfrog and get ahead? Or is the fact of open source a dangerous feature and instead the aim should be to reduce the number of actors in this field and to close down access and thus risks?

This conference considered three broad areas of likely effects of the creative destruction brought by rapid advances in AI: the impact on information systems; the future of work; and the impact on the relationship between government and the private sector, especially given the overpowering dominance of private corporations.

The impact on information systems

Disinformation and the spreading of fake news are clearly harmful as is the cherry-picking and selective promotion of information. LLMs are going to further automate the biases already in play, potentially further exacerbating social and political polarisation. AI-enabled visual disinformation (such as deep fakes) adds to this scenario, especially as the tools to detect it are not fully developed. Previously complicated tasks of editing videos and images have been simplified and enabled by AI. What measures are available to manage this content? Is there for example a means by which users could be alerted to synthetically generated content from generative AI models, to support judgements of their veracity? Ideas about labelling, watermarks, disclaimers for AI-generated content seemed weak in the face of the likely tsunami of content.

Who then is liable for the spread of false information? Can product liability be enacted against the technology companies? How far is it up to users to become proficient in critically assessing their own uses? But AI produces content that is well presented in an almost human-like format. The potential to use human voices (known to individuals) is qualitatively new and it is likely that we are (for the moment) more susceptible to believing this kind of content. At what point does it become the responsibility of democratic governments to maintain the integrity of information systems?

Can regulation be adapted to meet certain use case requirements? The aim would be to evaluate classes of use cases and use existing regulation. If we treat the models as aggregators of already existing content, then they would be affected by copyright laws. If we take its content to be a new product, then the liability could fall on the companies themselves, for example for spreading disinformation. Can *Section 230* protect a tech company from liability for third-party content? Many businesses depend on copyright laws to protect original work and to incentivise innovation. AI LLMs will disrupt some areas of copyright and create new grey areas over ownership of AI-generated

content. Is it possible to create an economic incentive to implement copyright laws and apply the three Cs: consent, credit and compensation? There was division between those who considered that existing regulation could be adapted to manage generative AI and those who thought existing regulation to be wholly inadequate.

Many of these models have been trained with online data garnered from people without their permission for these uses. There is a question of legitimacy and whether, for example, data drawn from social media apps is being 'permissioned' in the right way. There are also risks of false information being fed into the software.

On the plus side, was a potential for AI dashboards to support individual interests and decision-making. Forms of personalisation could allow citizens to regain control of their data and retake the power from companies, and could help in everyday tasks, even assisting in forms of personal development. But hyper-personalisation at a state- or larger level contains many other risks associated with the driving of biases and social fragmentation.

Impact on Work

Work is currently in a state of flux, and this will be accelerated by machine learning and LLMs. There will be far reaching impacts on paid (and unpaid) work. The short-term impact is likely to be in lower-skilled work. Jobs making boiler-plate text or functional graphics, often work that is outsourced to developing countries, will be lost. The UK economy was described as currently under-automated and the costs of deployment of chatbots in areas of white-collar jobs is low, creating much scope for automation.

There was a clear expectation too that AI could replace certain kinds of tasks that are *components* of existing jobs and perhaps even take out the drudgery by summarising, conducting literature searches, case comparisons, recommendation-based tasks and further automating the composition and replies to various communications. If so, this could save time for the work that only humans can do and increase human efficiency. Even in areas of creative thinking such as brainstorming, LLM's are a significant resource. Idea generation can also be somewhat automated, and the skills to use models shift to being able to effectively frame the questions. Overall, AI-powered tools could enable workers to tackle jobs previously out of reach. But many consequences are unknown. The overarching question now is how to speed up the equitable rollout of AI in the workplace because that will be necessary for companies to be competitive and therefore survive, while mitigating the negative effects of the transition.

It was also expected that tools such as ChatGPT will be exceptional aids for learning. Although a current pre-occupation is the use by pupils or students to cheat, there are many ways that they can be brought into the classroom to transform education systems.

Government and private sector

How should governments facilitate greater governmental expertise at state level and facilities to assess the geopolitical risks that come with this technological advance? A greater understanding of the available resource was said to be critical. To train leading-edge models, countries require cutting-edge GPU chips. Without them, there is little possibility to advance state-of-the-art technology. The capacity to understand the global flow of these components must be a part of government understanding of geopolitical security risks. Additionally, diversifying supply in compute capabilities is in the interest of the West. That said, a focus on resource and compute does not address harms that arise from applications.

So how can government and auditors get to grips with outcomes of generative AI based on LLMs? The global regulatory landscape is complex. A definitive, fixed form of law was thought impossible. A more likely means to address regulation could be via audits at three different levels: technology providers, models and applications. Would it be possible to build trust in regulation by developing expertise in particular domains (i.e. vertical regulation), rather than from a general perspective (horizontal regulation)? Could the scale of tech companies be better controlled and what safeguards are needed to safeguard vulnerable groups and to control new products?

Public-private partnerships were considered essential. For government there must be engagement with a wider range of leaders in industry, universities and the public sector. This will not only help utilise these technologies to improve public service delivery but will also enable more flexible and sensitive regulations on generational models in a way that better understands the issues involved.

With some concentration of AI talent in London, the UK was said to be in a pivotal, but rapidly closing, moment of being able to input ideas, alongside the powers in Silicon Valley, about the future of this industry. But to do so, governments need to better understand the trajectory of development over the next few years and the potential creation of synthetic AI (based on simulated data), which could further supplant some human capabilities. Together with the socio-human implications, governments should also recognise the economic potential that could result from a positive framing for generative AI outcomes. Additionally, for governments, there are the military implications, given the increasing dual use of these tech functionalities.

A first step would be to map out risks and opportunities. Forms of taxonomy exist within academia, but the public sphere does not yet have an accessible form of assessment to support public understanding of the social, economic and military risks and opportunities or the existential threats.

The UK cannot realistically initiate a technological race for AI and expect to win. However, it may find greater power in supporting a multilateral effort, leveraging its current advantage in AI as a bargaining chip for continued engagement (especially with the EU, for example). Additionally, given the precarious state of the world, state measures should not be reduced to simple protectionism. This is a new human issue and requires a global approach and global fora.

There are clear cultural and geographical differences that will play out – different countries are likely to use different training data based on their social, political and cultural views. For example, western countries may use the whole world wide web, whereas Chinese language models are only using a subset of data. Certain political systems will want significant, top-down influence to ensure that adaptation within/across models won't occur, effectively resulting in border controls on the internet.

These adaptations underscore the biases that can be baked in, depending on the cultural norms of populations interacting with the systems. A Korean application, 'Clova Call Service', demonstrated (for the conference) a system of automated phone calls to elderly citizens to check on their welfare. The demonstration of Clova AI's use of large language model technology showed the synthesis of speech and recognition + medical data + memory in an application which provides a service to contact and engage with elderly people, for example over their medical care or well-being. The service can draw on previous personalised conversations, update on medical conditions and recommend action to be taken and was described as accepted with Korean society.

Additional questions this conference did not address.

Global equity. Could there be an opportunity for leapfrogging or is it likely that inequalities will only be widened further? Data problems, labour and maturity of tech locally might limit potential in these regions.

Security and geopolitical questions were not discussed enough – security in leaving data audit trails and in terms of the powerful capabilities of chatbot, plus memory and in the possible development of new tech borders or AI models for different political jurisdictions. Will there be divergence between western and Chinese models, which in turn could have impacts on populations?

The impact on climate change trends and the environmental footprint of AI model supply chains and energy use was raised but not considered.

The potential for embodiment. The capabilities that can arise from large language models plus API plus physical output such as robots or drones were pointed to but not followed up.

Societies have incorporated technology many times over and it is likely that this form of AI will not be the last new tech human societies will have to absorb. Can societies get better, even good at this? Or should the speed of the development of this tech and the possible link to AGI be slowed? The dominance of the private sector is overwhelming: 99% of this tech and the massive investment is being driven by the private sector. Access to the major private labs (and to the necessary compute resource) is limited. There is no equivalence for AI safety in terms of biosafety labs, genetic engineering controls, anthrax etc. What happens if AI were to run communication functions within companies, or otherwise infiltrate company management?

This Note reflects the Director's personal impressions of the conference. No participant is in any way committed to its content or expression.

PARTICIPANTS

AUSTRALIA

Ms Cass Matthews
Office of Responsible AI, Microsoft.

BELGIUM

Mr Nicolas Moës
Director for European AI Governance, The Future Society.

CANADA

Mr Serge Blais M.Sc. (A), Exec MBA
Executive Director, Professional Development Institute, University of Ottawa.

Ms Rebecca Finlay
CEO, Partnership on AI.

Mr Marc-Etienne Ouimette
Global lead for AI policy, AWS (Amazon Web Services).

Mr Nitarshan Rajkumar
PhD Candidate in Artificial Intelligence, University of Cambridge.

CANADA/UK

Lady Rosemary Leith Berners-Lee
Venture Investor; Fellow, Berkman Center, Harvard University; Co-Founder, World Wide Web Foundation.

JAPAN

Mr Ren Ito
COO, Stability AI.

REPUBLIC OF KOREA

Jung-Woo Ha PhD
Head, NAVER AI Lab, Seoul.

SOUTH AFRICA

Mr Muhammed Razzak
Rhodes Scholar; DPhil Candidate in Computer Science, University of Oxford; Student Researcher, Alan Turing Institute.

UNITED KINGDOM

Mr John Bachelor
Economic Policy Advisor (and lead on AI policy), Labour Party.

Sir Tim Berners-Lee
Inventor of the World Wide Web; Founder and Director, World Wide Web Consortium and the Web Foundation; co-founder and President, Open Data Institute, London.

Mr Blake Bower
Director, Digital and Technology Policy Directorate, Department for Science, Innovation and Technology.

Mr Matt Clifford MBE
Co-Founder and Chief Executive, Entrepreneur First, London.

Mr Alex Creswell OBE
Co-chair, Turing Innovation Hub (Manchester); Chair in AI and Digital, Manchester University. SVP Public Policy, Graphcore Ltd.

Mr Patrick Gilday
Angel investor and advisor to several leading Machine Learning companies across the UK.

Miss Sophie Hackford
Futurist and advisor to John Deere & Co and New Lab, Brooklyn.

Mr Ian Hogarth
Co-founding partner, Plural; Chair, Phasecraft, a leading quantum computing start-up.

Ms Hannah Rose Kirk
DPhil Candidate in Social Data Science, University of Oxford; researcher, Online Safety team, The Alan Turing Institute.

Dr Pushmeet Kohli
Head of AI for Science, DeepMind.

Ms Christina Last
US-UK Fulbright scholar and postgraduate student at MIT; CTO, AQAI.

Mr James Lawson
Senior Special Adviser in the Cabinet Office; Senior Fellow, Adam Smith Institute.

Professor Dame Angela McLean DBE, FRS
Government Chief Scientific Adviser.

Mr Rajay Naik
Chief Executive Officer, Skilled Education; Chairman, UK Commission on Lifelong Learning. A Governor and member of the Programme Committee, The Ditchley Foundation.

Mr Tom Nixon
Director, Faculty Science Ltd.

Professor Sir Nigel Shadbolt FRS FREng
A leading researcher in Artificial Intelligence; Principal, Jesus College, University of Oxford and a Professor of Computing Science, University of Oxford; Chairman, Open Data Institute.

UK/USA

Mr Jack Clark
Co-founder, Anthropic.

Ms Kay Firth-Butterfield LLM, MA
Head of Artificial Intelligence and a member of the Executive Committee, World Economic Forum.

Professor John Tasioulas
Professor & Director, Institute of Ethics in AI, University of Oxford.

Mr Matt Warman MP
Member of Parliament (Conservative) for Boston and Skegness (2015-); former Minister for Digital Infrastructure, Department for Digital, Culture, Media and Sport,

Dr Michael Webb
Former Chief Economic Adviser to Prime Minister and Chancellor.

Professor Michael Wooldridge
Head, Department of Computer Science, University of Oxford.

UNITED STATES OF AMERICA

Ms Julie Brill
Corporate Vice President and Chief Privacy Officer, Microsoft Corporation.

Mr Cristian Canton PhD
Director of research and engineering, Responsible AI division, Meta.

Mr Kenneth Cukier
Deputy Executive Editor, correspondent and editor, The Economist.

Professor Yang-Hui He
Fellow, London Institute; Professor of Mathematics, City, University of London.

Dr Alfred Z. Spector
Visiting Scholar, Massachusetts Institute of Technology.

PLUS

Mr Azeem Azhar
Author and producer, Exponential View newsletter and podcast.

Dr Nathan Benaich
Founder and General Partner, Air Street Capital.

Mrs Kata Escott
Head, Office for Science and Technology Strategy, Cabinet Office.

Mr Connor Leahy
CEO and co-founder, Conjecture.