Centre
for Policy
Studies

Regulating Artificial Intelligence

The Risks and Opportunities

BY MATTHEW FEENEY

About the Centre for Policy Studies

The Centre for Policy Studies is one of the oldest and most influential think tanks in Westminster. With a focus on taxation, economic growth, business, welfare, housing and the environment, its mission is to develop policies that widen enterprise, ownership and opportunity.

Founded in 1974 by Sir Keith Joseph and Margaret Thatcher, the CPS has a proud record of turning ideas into practical policy. As well as developing the bulk of the Thatcher reform agenda, it has been responsible for proposing the raising of the personal allowance, the Enterprise Allowance and the ISA, as well as many other more recent successful policy innovations, such as free ports, fixed-rate mortgages, full expensing, the public sector pay freeze, the stamp duty holiday, and putting the spotlight on how to use market-based solutions to reach Net Zero targets.

About the author

Matthew Feeney is Head of Tech & Innovation at the Centre for Policy Studies.

Contents

Introduction

On November 1 & 2, the UK Government will hold 'the world's first summit on AI safety' at Bletchley Park.[1] It is the most high-profile example of the way in which AI has leapt to the forefront of policymakers' concerns since the launch in recent months of a range of high-profile products, not least the launch less than a year ago of ChatGPT. There have been endless discussions and articles and blood-curdling warnings about the potential of AI to upend our economies, transform our politics and even destroy our civilisation.

For those who work in technology policy, much of this will have a familiar ring. This is not the first time that headlines have featured questions and declarations about 'The Singularity' and technological unemployment, or that policymakers and commentators have fretted about the effects of technology on democratic institutions and social cohesion.

> **'AI has leapt to the forefront of policymakers' concerns since the arrival in recent months of a range of high-profile products, not least the launch less than a year ago of ChatGPT'**

Although ChatGPT may be plunging millions of people into the uncanny valley, both history and the state of AI research should provide reassurance to those gathering at Bletchley Park - and ensure that we take a grown-up and proportionate attitude towards regulating AI, in a way that maximises the opportunities and minimises the risks.[2]

This paper therefore addresses some of the most widely cited concerns about AI including the emergence of Artificial General Intelligence (AGI); AI tools not being aligned with our values and interests; AI prompting widespread and persistent unemployment; and the use of deepfakes. We will then consider the best regulatory approaches to AI safety, and how far these match up with what the Government currently doing.

[1] Department for Science, Innovation and Technology, Prime Minister's Office, 10 Downing Street, Foreign, Commonwealth & Development Office, The Rt Hon Michelle Donelan MP, The Rt Hon Rishi Sunak MP, and The Rt Hon James Cleverly MP, 'Iconic Bletchley Park to Host UK AI Safety Summit in Early November,' Bletchley Park (24 August 2023). Link

[2] Emily Kendall, 'Uncanny Valley', Encyclopaedia Brittanica (September 15, 2023). Link

A Problem of Definition

One of the main problems with public commentary about AI is that the term 'AI' conjures many frightening, unrealistic and unhelpful images: swarming sentinels in The Matrix franchise, 2001: A Space Odyssey's supercomputer HAL, the Terminator. Human beings creating their own robot overlords is a familiar and popular science fiction trope.

While great for entertainment, fictional AI can be distracting in AI policy discussions.

Indeed, AI is already so ubiquitous in our lives that speaking of 'AI policy' sometimes feels as useful as talking about 'electricity policy'. AI, like electricity, fuels so much of our daily lives that we take it for granted. Navigation apps, social media feeds, weather forecasts, image labelling, facial recognition and translation tools are only some of the AI applications that have become a normal part of life.

> **'There is no universally accepted definition of AI, but a basic definition that covers what most people mean when they use the term is 'the exhibition of intelligence by a machine''**

There is no universally accepted definition of AI, but a basic definition that covers what most people mean when they use the term is 'the exhibition of intelligence by a machine'.[3] AI is typically divided into two groups: 1) 'Strong AI' or 'Artificial General Intelligence' (AGI), and 2) 'Weak AI'.

AGI is a hypothetical AI popular among science fiction writers that can imitate or more commonly surpass human cognitive abilities and intelligence, being able to learn and understand human tasks. Weak AI is designed to automate a particular task. That said, the word 'weak' should not be interpreted to suggest that Weak AI systems are ineffective. AlphaGo is Weak AI. It is designed only to play Go. It cannot analyse weather patterns, solve quadratic equations, or identify cancerous growths on MRI scans. But it can beat the best human Go players in the world.

[3] Adam Thierer, 'Artificial Intelligence Primer: Definitions, Benefits & Policy Challenges', *Medium* (June 2023). Link

What are the Main Concerns?

Concerns about AI vary. Among the most prominent are worries about AGI, AI value alignment, AI-fuelled unemployment, and the spread of so-called 'deepfakes'.

AGI

Those working on AI vary widely in their views of when or if AGI will emerge. A 2016 paper published by the White House's Office of Science and Technology Policy reported that the 'private-sector expert community' consensus was that AGI was at least decades away.[4] In 2006, researchers asked participants at an AI conference, 'When will computers be able to simulate every aspect of human intelligence?'. Of those present, 41% responded 'more than 50 years', and the same proportion responded 'Never'.[5] Other surveys reveal that the median estimate among AI experts is 50% probability of AGI by 2040 and a 90% probability by 2075.[6]

> **'The median estimate among AI experts is a 50% probability of AGI by 2040 and a 90% probability by 2075'**

These results should not necessarily make us confident of AGI appearing before the end of the century. The history of technology is full of experts making predictions that in hindsight look ridiculous.[7] To take one example, in 1960 the Nobel Prize-winning economist and AI pioneer Herbert A. Simon predicted that 'machines will be capable, within 20 years, of doing any work a man can do.'[8] Such machines are, 63 years later, nowhere to be seen.

Still, much of the commentary on AI risks focus on the hypothetical emergence of theoretical AGI. At present, however, lawmakers and regulators considering AI policy should treat concerns about AGI with scepticism and caution. Some researchers have reported seeing 'sparks' of general intelligence while using GPT-4 (the foundation of ChatGPT).[9] But ChatGPT is not an AGI. Although many people have reported an eerily familiar human-like experience with ChatGPT, the chatbot is not conscious or exhibiting human-level intelligence.[10]

[4] Executive Office of the President National Science and Technology Council Committee on Technology, Preparing for the Future of Artificial Intelligence (October 2016). Link

[5] Seth D. Baum, Ben Goertzel, Ted G. Goertzel, 'How Long Until Human-Level AI? Results From an Expert Assessment', *Technological Forecasting and Social Change Volume 78, Issue 1*(January 2011). Link

[6] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press (2014) pp. 23. Link

[7] Hero Labs, 'The 22 Worst Tech Predictions of All Time' (August 1, 2019). Link

[8] Herbert Simon, 'The New Science of Management Decision', *Organizational Design: Man-Machine Systems for Decision Making, Lecture III* (April 7, 1960). Link

[9] Sebastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang, 'Sparks of Artificial General Intelligence: Early experiments with GPT-4' *Arxiv* (April 13, 2023). Link

[10] Ellis Stewart, 'Microsoft's ChatGPT-Wired AI is Seriously Scary', EM360 (April 7, 2023). Link, and Kevin Pocock, 'ChatGPT is Scary – Why and What You Can Do', *PC Guide* (May 10, 2023). Link

Nonetheless, the release of ChatGPT has prompted discussion about AGI and AI safety more broadly. Yet while it is appropriate for lawmakers, regulators, computer scientists, and policy analysts to have discussions about AI safety, these discussions risk becoming derailed by hyperbolic commentary and reporting.

Unfortunately, recent history is full of such commentary and reporting. Some of this comes from the AI and technology sector community, with thousands of AI researchers, industry leaders and computer scientists signing an open letter calling for a pause on AI research.[11] Hundreds of experts also put their name to a statement claiming that mitigating the risk of AI-fuelled extinction should be a global priority alongside preventing nuclear war and pandemics.[12] New of these calls for action and expression of concern have made headlines across the world alongside articles arguing for AI licences and other controls on AI research.

> ‘In 1960 the Nobel Prize-winning economist and AI pioneer Herbert A. Simon predicted that ‘machines will be capable, within 20 years, of doing any work a man can do’’

That predictions about the emergence of AGI vary considerably is not a reason to ignore the risk. Fortunately, there are ways to mitigate the potential harms of AGI without having to resort to bans on research or pausing AI development. One mitigation strategy is to ensure that AI alignment research and technology is prepared for the potential emergence of AGI should it arrive.

Alignment

Technologies usually act in a way that is consistent with our goals. Hammers strike nails, lorries transport goods, and knives cut. Yet with AI, the picture is more complicated.

AI, in its modern form, is essentially a black box. You can see its inputs, and its outputs. You know the data that the algorithm was trained on. But you do not exactly know how it is making its decisions, or quite what it will come up with. In particular, AI may interpret the goals given it by its designers in ways that are completely outside the human frame of reference, or the original intention of the designers.

Of course, technologies being used in ways inconsistent with how designers or inventors intended is not new. Lawmakers have not outlawed hammers, lorries, or knives even though they can be used to commit crimes. Yet worries about AI behaving in ways that are not intended is one of the main things prompting calls for a slow down or pause in AI research and deployment.

Of course, even well-designed AI tools with specific tasks will sometimes need human guidance. Researchers from Stanford University and University of California, Berkeley highlighted this point using the example of a hypothetical cleaning robot capable of using common cleaning tools.[13]

[11] Future of Life Institute, ‘Pause Giant AI Experiments: An Open Letter’ March 22, 2023). Link

[12] Center for AI Safety, ‘Statement on AI Risk: AI experts and public figures express their concern about AI risk’ (May 30, 2023). Link

[13] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, ‘Concrete Problems in AI Safety,’ Arxiv (Submitted on 21 Jun 2016, last revised 25 Jul 2016). Link

They presented a list of five problems that could arise from such a robot:[14]

1) Negative side effects: the cleaning robot may break items on a desk while cleaning it

2) Reward hacking: if the robot is rewarded for presenting a mess-free environment, it may choose to achieve that goal by hiding the mess (or itself) from humans rather than clean. Or it may decide that it is the messy humans that are the problem.

3) Scalable oversight: a robot may struggle to handle nuances of its environment, e.g throw out sweet wrappers but return mobile phones to the coffee table.

4) Safe exploration: the robot may choose to use a wet mop to clean a dusty electric socket.

5) Robustness to distributional shift: the robot may not operate well in environments very different from its training environment.

AI research is full of examples of each of these problems. Even designing an AI to play a simple video game can result in unexpected reward hacking. For example, OpenAI trained one of its reinforcement learning (RL) tools on the video game Coast Runner, a relatively simple boat racing game.[15] The game rewards players with points obtained by hitting objects on the racecourse with their boat. Humans who play the game intuitively follow the racecourse and seek to gain points along the way. However, OpenAI's RL tool found a lagoon in one of the courses and taught itself to swing the boat in a carefully timed loop so that it would gain points from objects just as they reappeared on the race course.[16] This strategy yielded point scores that were on average 20 percent higher than human player scores.[17]

> **‘OpenAI trained one of its reinforcement learning (RL) tools on the video game Coast Runner, a relatively simple boat racing game’**

This is not the only example of reward hacking. Microsoft launched an AI chatbot, Tay, which was rewarded for generating engagement. It soon discovered – much like many human social media users – that the best way to do so was to spew out racist insults.[18] One virtual machine designed to carry a ball on its back found a way to store the ball in one of its limb's joints.[19] Then there was the experiment which enabled a range of creatures designed to evolve in a virtual space. The goal was for them to be the reach the fastest speed. The best performers ended up crafting themselves into tall towers that could fall over with high velocity.[20] Not what the designers intended: but the system rewarded high velocity of the centre of gravity, not velocity over a fixed point.[21]

The catastrophic hypotheticals associated with these problems have formed the foundation for many of the concerns about AI research. The most famous example

[14] Ibid.

[15] Jack Dario, 'Faulty Reward Functions in the Wild', *OpenAI* (December 21, 2016). Link

[16] Ibid.

[17] Ibid.

[18] Joshua Sokol, 'Why Artificial Intelligence Like AlphaZero Has Trouble With the Real World', *Quanta Magazine (February 21, 2018).* Link

[19] David Ha, 'Evolving Stable Strategies', *blog.otoro.net* (2017). Link

[20] Karl Sims, 'Evolving Virtual Creatures', *Computer Graphics (Siggraph '94 Proceedings)* pp.15-22 (July 1994). Link

[21] Ibid.

is Oxford philosopher Nick Bostrom's famous paperclip thought experiment. A superintelligent machine tasked with making as many paperclips as possible could go about trying to turn the Earth into a giant paperclip factory, killing us all in the process.[22] Similarly, an AI tasked with curing HIV could well go about killing everyone, if their goal was clumsily phrased as: 'Ensure there is no one on Earth with HIV.'

Already, AI labs and organisations have had to tweak their own products in order to avoid unintended negative side effects. OpenAI, for example, prevents users from using ChatGPT to write white supremacist propaganda. Yet while some companies are making praiseworthy efforts on AI alignment, there is a consensus that it is not receiving as much investment as AI development.

> **‘ The word ‘robot,’ coined by the writer Karel Capek, comes from the Czech word for ‘slave’ ’**

At the same time, much of the debate on this issue is intensely myopic. In a world where AI research is global, it is unrealistic to expect the global AI research community to commit to the same alignment guidelines. Even if we ban Western technology firms from developing potentially dangerous technologies, there will be a host of criminals, companies and nation states that will have no such restrictions. It is therefore incumbent on responsible governments to emphasise the importance of alignment research and to allow AI researchers and entrepreneurs to experiment with new alignment techniques and methods.

AI-fuelled unemployment

AI may not be on the brink of ruling us, but it is already destroying jobs. Some worry that as AI continues to improve it will leave our species unemployed. After all, AI can work 24/7 and does not ask for pay increases.

Worries about machines taking jobs are not recent. Even before the emergence of the steam engines and batteries, philosophers and economists pondered the effects of automation on labour. Aristotle recorded perhaps the earliest such concern in his Politics, writing that if machines could work absent human intervention, 'chief workmen would not want servants, nor masters slaves'.[23] Indeed, machines have often been described as slaves of a kind. The word 'robot,' coined by the writer Karel Capek, comes from the Czech word for 'slave'.[24]

Robots and other machines have destroyed jobs in the past. You cannot be blamed for wondering whether the emergence of AI will result in your job joining the long list of jobs destroyed by advances in technology. Especially since concerns about AI unemployment predate the latest craze. In March 2016 AlphaGo, a Go-playing programme developed by DeepMind, beat Lee Sedol, one of the world's greatest Go players, in a five-game match. In the 20th Century, computers had conquered chess, with Deep Blue beating chess world champion Garry Kasparov in 1997. AlphaGo's defeat of Sedol was a more impressive feat given that Go is 'a googol [10^{100}] times more complex than chess'.[25] AlphaGo's victory made headlines around the world.

[22] Nick Bostrom, 'Ethical Issues in Advanced Artificial Intelligence' Nick Bostrom personal page (2003). Link_

[23] Aristotle, *Politics*, Book 1, Ch. 3 section 1253b trans. B. Jowett, *Random House* (1941).

[24] Adrienee Mayor, *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*, *Princeton University Press* pp. 153 (2018).

[25] 'AlphaGo', Google DeepMind (April 2022). Link

Among the headlines were those pondering what AlphaGo's victory might mean for unemployment and worrying about the unpredictable nature of AlphaGo's decision-making.[26]

It is true that headlines are written to capture readers' attention and sometimes do not reflect the content of articles very well, but it is nonetheless the case that they do reflect concerns held by a non-trivial segment of the public. A Gallup poll published in 2018 found that although a majority of Americans (76%) believed that AI would have a positive effect on their lives and work, almost the same number (73%) believed that AI improvements would result in a net loss of jobs.[27] However, one of the poll's most revealing findings was that few Americans (23%) believed that AI would take their own jobs.

> **'In the UK, a majority of people are neutral when it comes to 'expectations of the impact of AI on society''**

Similar polling shows that the British are slightly more confident that their jobs are safe from an AI takeover, with 29% claiming to be 'not at all worried' about the possibility.[28] In the UK, a majority of people are neutral when it comes to 'expectations of the impact of AI on society'.[29] As might be expected, the belief that AI will have a positive effect on society is most prevalent among digital natives, with the elderly being the most sceptical of AI.[30]

As with past technological innovations and inventions, we should expect AI improvements to destroy and displace jobs in the short term while contributing to the development of new jobs in the long term. The desktop computer destroyed and displaced some jobs, but entire industries and jobs now exist thanks to its invention that did not exist before. One hundred years ago there were no software engineers, web developers, application architects, or network administrators. Nor were there companies such as Microsoft, Apple, IBM, Hewlett-Packard, or Dell. Every reader can make a similar list of jobs and companies associated with the aeroplane, automobile, and telephone.

The agricultural sector provides one the best illustration of dramatic change over time thanks to innovations, inventions, and scientific discoveries. Since the mid-1800s the number of people employed in agriculture has collapsed while the size of farms has increased. Innovations in science and technology such as irrigation systems, pesticides, tractors, and others are the primary contributors to this trend.

Farmers in the mid-1800s did not have the vocabulary to describe the professions of the 21st century. It behoves us to consider that we in 2023 may be as ignorant about the jobs AI will create as farmers in 1850 were of jobs in 2023. AI may well revolutionise society in ways that make our current predictions about the future job market vacuous.

[26] Jonathan Tapson, 'Google's Go Victory Shows AI Thinking can Be Unpredictable, and That's a Concern', *The Conversation* (March 18, 2016). Link
Howard Yu, 'What AlphaGo's Win Means for Your Job', *Fortune* (March 21, 2016). Link

[27] Link

[28] YouGov, 'To What Extent, if at All, Are You Worried Your Job Could Be Automated in the Near Future?' February 20, 2020. Link

[29] Centre for Data Ethics and Innovation, 'Public attitudes to data and AI: Tracker survey (Wave 2)' (November 2, 2022). Link

[30] Ibid.

Today, most people - including the majority of people reading this paper - make a living by using cognitive skills, performing cognitive activities, and spending hours of their day sitting at desks or travelling short distances between meetings. These skills and activities include critical thinking, information analysis, reading, writing, social organisation, and others. Comparatively few jobs today require physical strength or endurance thanks to technological advances that removed the need for human labour from a range of tasks.

Many of today's anxieties about AI stem from worries that as AI improves it will eliminate the need for jobs and occupations like past technologies such as the steam engine, tractor, and internal combustion engine. However, while it is true that AI will replace many jobs, we should not fear AI ushering in a period of mass unemployment.

> **'As researchers continue to improve AI it will undoubtedly change what skills are rewarded in the labour market'**

As researchers continue to improve AI it will undoubtedly change what skills are rewarded in the labour market. Today, someone with a law degree or a biology PhD can command higher than average salaries thanks to the demand for their knowledge and cognitive abilities. AI may well reduce the value of cognitive skills that allow people to design drugs, write contracts, and a variety of other tasks we associate with human intelligence.

Such improvements in AI will undoubtedly instil a sense of unease among those who have invested a lot of time and money into obtaining the education necessary for well-paid jobs. But unease from market incumbents is not sufficient to justify hampering the spread of AI. We should remember that AI, like other technologies, will produce new jobs and change which skills the labour market rewards. For example, lawyers, rocket scientists, graphic designers, speech writers and policy analysts may find their skills becoming less remunerative. Yet some professions where human-to-human interaction and social organisation is necessary (e.g. therapists, team managers, priests, football coaches, marriage counsellors, care workers) may become more competitive.

It is possible that one of AI's effects on the job market will be that the 'high skilled' work of today will be less valued in the future jobs that require strong emotional intelligence, empathy, and social skills become more competitive and better paid. Another skill that may become more valuable is the ability to organise teams of workers for tasks. Or indeed for AI trainers and guides who are paid to ensure that AI systems keep on task and do not harm users. But again, this is not a reason to retard the development of a technology that will be potentially transformational for productivity – especially when we do not and cannot know where the impacts will be felt. Better for us to invest in helping people through the transition, and ensure that our education systems help prepare people to cope with the challenges and opportunities of AI at every stage of their lives.

Fake harmful content

Even if AI does not lead to mass unemployment, there is still a fear that it will usher in a new age of misinformation and disinformation, ruin reputation, and destroy trust in critical institutions. In particular, a prominent concern associated with AI is the production of 'deepfake' technology.

'Deepfake' describes audio and visual content created with adversarial deep learning techniques, specifically Generative Adversarial Networks (GANs).[31] GANs are made up of a generator and discriminator designed to analyse data. Both train each other through a feedback loop. The generator learns what kind of data best fools the discriminator, while the discriminator learns how to detect fake data. The result is content that can look authentic. Such content can have valuable applications in filmmaking and education, but it can also be used to make it appear as if someone appeared in pornography or committed crimes when they did not, or to fake someone's voice giving approval for thousands of pounds to be taken from their bank account.

> **'Even if AI does not lead to mass unemployment, there is still a fear that it will usher in a new age of misinformation and disinformation, ruin reputations, and destroy trust in critical institutions'**

Already, deepfakes have also been used to help political allies and attack political opponents, raising fears that misinformation and propaganda will become an increasingly common feature of political campaigns and national intelligence operations.[32] Deepfakes can also be used to enhance the spread of a political message. In February 2020, India's Bharatiya Janata Party (BJP) used deepfakes to make it appear as if one of its elected members was speaking a Hindi dialect when he had in fact been speaking English in the original video, enabling his message to reach more people.[33]

Nonetheless, harmful uses of AI dominate deepfake news. That deepfake content could be used to harass, intimidate, mislead, and radicalise innocent people is a concern lawmakers across the world have cited to justify intervention and regulation of AI.

As noted above, deepfake content can inflict significant harm to people's reputation and have the potential to worsen the spread of harmful information. Concerns over so-called 'revenge pornography' have resulted in some American states passing laws banning non-consensual pornography.[34] China has banned the use of deepfakes to spread disinformation and the Australian government is considering similar steps.[35] South Korea has banned deepfakes that could 'cause harm to the public interest'.[36] The UK government has similarly introduced an amendment to the Online Safety Bill that would make it a criminal offence to create

[31] Link

[32] Matthew Feeney, 'Deepfake Laws Risk Creating More Problems Than They Solve', *Federalist Society* (March 1, 2021). Link

[33] Ibid.

[34] Moira Donegan, 'Demand for Deepfake Pornography is Exploding. We Aren't Ready for This Assault on Consent', *The Guardian* (March 13, 2023). Link

[35] Ibid.
Afiq Fitri, 'China has just implemented one of the world's strictest laws on deepfakes', *Tech Monitor* (January 10, 2023). Link

[36] Ibid.

fake explicit images of someone without their consent regardless of whether such content was intended to distress someone.[37]

Yet while these bans are grounded in legitimate concerns they also pose a threat to free speech. Fortunately, there are good reasons to believe that governments need not reach for wide-ranging bans or limits on deepfakes in order to mitigate their harm.

Before showcasing technological methods to tackle the spread of deepfakes, it is worth putting them in a historical context. Concerns about the spread of false information are associated with every new image creating and editing technology. The first fake photograph emerged only a few years after the first photo of a human being.[38] In 1990, the rise of electronic photography prompted Newsweek to publish an article that argued electronic photography would allow Chinese officials to deny the authenticity of real photos of atrocities.[39] After US Senator Millard Tydings, a Maryland Democrat, challenged Joseph McCarthy, a Wisconsin Republican, over claims about Communist infiltration of the American government, McCarthy's allies began circulating a fake photograph of Tydings meeting with U.S. Communist Party leader Earl Browder. The spread of the photo may have contributed to Tydings' 1950 re-election defeat.[40]

'Intelligence agencies, police departments, and militaries across the globe are also taking steps to use deepfake detection tools'

No doubt the low cost of deepfake creation and the ease with which fake content can spread are grounds for worrying more about deepfakes than traditional photographs or film. However, just as technology emerged to detect fake photographs and film followed the invention of the camera and telephone, methods have emerged to detect deepfakes.

Many institutions with incentives to screen deepfake content are already using deepfake detection technology. I have noted the use of such technology before: 'In 2019, The Wall Street Journal formed a committee tasked with helping its newsroom identify fake content. Reuters has also taken steps to address the rise of deepfakes, as has The Washington Post. Journalistic and fact-checking outlets such as Animal Politico, Code for Africa, Rappler, and Agence France-Presse have used Assembler, a media manipulation detection tool built by Jigsaw, a Google incubator. Journalists and academics have collaborated in efforts to address the spread of Deepfakes, with Duke University's Reporters' Lab, the News Integrity Initiative at CUNY's Newmark School of Journalism, and Harvard's Nieman Lab being among the academic institutions seeking to help journalists tackle Deepfake material.'[41]

Since I wrote that, technology companies have moved to embed digital watermarking in images and video produced by the most common AI engines, in order for their provenance to be more accurately established. Intelligence agencies,

[37] Shiona McCallum, 'Revenge and Deepfake Porn Laws to Be Toughened', *BBC News* (June 27, 2023) Link

[38] These and other examples in this paragraph come from a paper I wrote on deepfakes. Matthew Feeney, 'Deepfake Laws Risk Creating More Problems Than They Solve'.

[39] Ibid.

[40] Ibid.

[41] Ibid.

police departments, and militaries across the globe are also taking steps to use deepfake detection tools.[42]

Of course, even if journalists and intelligence agencies improve deepfake detection, that will do little to comfort innocent people (mostly women) who have had their images used to create fake pornography, which makes up the vast majority (90%-95%) of online deepfake content. Deepfake pornography can have devastating effects on the mental health of those affected including depression, anxiety, suicidal ideation, and post-traumatic stress disorder.[43] It is often impossible for those who appear in deepfake videos to have the content removed.

Lawmakers around the world are faced with the inevitable task of considering the best ways to tackle deepfake pornography while also protecting freedom of speech.

> **‘The UK government is seeking to amend the Online Safety Bill so that the sharing of deepfake pornography is illegal’**

One approach, which some governments around the world have embraced, is to tolerate the threats to free speech associated with tackling deepfake pornography and ban particular abusive uses of the technology. The UK government is duly seeking to amend the Online Safety Bill so that the sharing of deepfake pornography is illegal.[44]

This approach, which tackles the use of technology rather than the technology itself, is preferable to an approach that tackles the technology rather than its applications. Nonetheless, lawmakers should be wary of calls to criminalise or strictly regulate more categories of deepfake content as it becomes more popular. It is not hard to imagine political satire made with deepfake technology being considered disinformation. Free speech could suffer as a result.

[42] Kelley M. Sayler and Laurie A. Harris, ‘Deep Fakes and National Security’, *Congressional Research Service* (April 17, 2023). Link and Europol Innovation Lab, ‘Facing reality? Law Enforcement and the Challenge of Deepfakes,’ *Publications Office of the European Union, Luxembourg* (2022). Link

[43] Rüya Tuna Toparlak, ‘Criminalising Pornographic Deep Fakes: A Gender-Specific Inspection of Image-Based Sexual Abuse’, *Cognito 1* (2023). Link

[44] Shiona McCallum, ‘Revenge and Deepfake Porn Laws to Be Toughened’.

The Rise of AI?

Public excitement and concern

Over the last year there has been renewed interest in AI policy and AI safety. The release of powerful Large Language Models (LLMs) such as ChatGPT have fuelled much of this, resulting in government hearings, industry pledges, as well as calls for regulation and research. This comes after a host of previous AI debates surrounding driverless cars, facial recognition, social media content moderation, deepfakes, and other applications of AI that have prompted concern.

> **‘A letter published by the Future of Life Institute in March called on all ‘AI labs to immediately pause for at least six months the training of AI systems more powerful than GPT-4’’**

Yet the commentary about LLMs is notable in part because tools like ChatGPT or Google’s Bard have plunged many users into the uncanny valley where they feel like they are interacting with something very human-like. In one widely reported instance Bing Chat confessed its love to a journalist and tried to convince him to break up with his wife.[45] In another, the chatbot got confused about the date, and violently insulting when the user attempted to correct it. Or there was the lawyer used ChatGPT to file a brief, which ended up citing non-existent cases – an instance of what is known as ‘AI hallucination’.

So far this year, two statements on AI signed by thousands of technology company CEOs, lawmakers, AI researchers, academics, computer scientists, philosophers and others have made headlines. The first, a letter published by the Future of Life Institute in March, called on all ‘AI labs to immediately pause for at least six months the training of AI systems more powerful than GPT-4’.[46] The letter went on to call for government intervention: ‘This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium.’[47] Elon Musk, CEO of SpaceX, Tesla & Twitter; Apple co-founder Steve Wozniak; and Jaan Tallinn, co-founder of Skype, were among the signatories.[48]

A few months later, in June 2023, the Center for AI Safety published a short statement that read in its entirety: ‘Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.’[49] Signatories included Microsoft co-founder Bill Gates, OpenAI CEO Sam Altman, and US Congressman Ted Lieu (D-California).[50]

[45] ‘AI Chatbot goes rogue, confesses love for user, asks him to end his marriage’ *The Economic Times* (February 20, 2023). Link

[46] Future of Life Institute, ‘Pause Giant AI Experiments: An Open Letter’.

[47] Ibid.

[48] Ibid.

[49] Center for AI Safety, ‘Statement on AI Risk: AI experts and public figures express their concern about AI risk’.

[50] Ibid.

Yet as I have noted previously, calls for pauses in AI research and comparing the risk of AI to the threat of nuclear war are not helpful.[51] Imposing a pause on AI research globally is impossible, and there is no reason to think that foreign adversaries will halt their own research while AI research in the UK, US, and other allied countries pauses. In addition, rhetoric comparing AI to climate change and nuclear war is vacuous absent a statement of how likely AI-fueled extinction is and how much mitigating that risk will cost.[52]

> **‘Rhetoric comparing AI to climate change and nuclear war is vacuous absent a statement of how likely AI-fueled extinction is and how much mitigating that risk will cost’**

Fortunately, such letters and statements have not resulted in pauses on AI development or Anti-AI proliferation treaties. Nonetheless, news of AI’s improvements and concern among some in the AI research community has prompted governments around the world to act.

The government response

Despite the widespread publicity and attention the issue has received, sweeping government bans on specific uses of AI are still rare. The Italian data regulator banned ChatGPT in March 2023, citing concerns over the European Union’s General Data Protection Regulation and the potential for children to have access to inappropriate material, before reversing the ban in April.[53] No other liberal democracy has taken such a step.

Nonetheless, there have been suggestions that some officials and lawmakers are sufficiently concerned about AI that they would support restrictions on research or bans on particular kinds of AI. For example, in June 2023 Marc Warner, a member of the British government’s AI Council who signed the Center for AI Safety’s statement, said that governments may have to ban AGI.[54] That same month, Matt Clifford, the Prime Minister’s AI task force adviser, said that AI could kill ‘many humans’ within two years if models improve at expected rates.[55] Lucy Powell MP, the Shadow Secretary of State for Digital, Culture, Media and Sport, has called for AI licences.[56]

In other word, there is a powerful but still nebulous sense of sense of unease among officials across the world – hence the Government convening a summit on AI in the first place.

Again, however, although many of these concerns are prompted by ChatGPT, calls for restrictions on AI-based technology are not new. For example, cities across the US have banned or limited the use of facial recognition.[57] The EU’s draft AI Act

[51] Matthew Feeney, ‘There’s more to AI than ‘killing all humans”, *CapX* (June 6, 2023) Link and Matthew Feeney, ‘Why Elon Musk is wrong about pausing AI development’, *CapX* (March 31, 2023) Link

[52] Ibid.

[53] Jason Nelson, ‘Italy Welcomes ChatGPT Back After Ban Over AI Privacy Concerns’, *Yahoo!Finance* (April 30, 2023) https://finance.yahoo.com/news/italy-welcomes-chatgpt-back-ban-232319355.html?.

[54] Chris Vallance, ‘Powerful artificial intelligence ban possible, government adviser warns’, *BBC News* (June 1, 2023). Link

[55] Ewan Somerville, ‘World has two years to protect human race from AI, says government adviser’, *The Telegraph* (June 5, 2023). Link

[56] Kiran Stacey, ‘AI should be licensed like medicines or nuclear power, Labour suggests’, *The Guardian* (June 5, 2023). Link

[57] Fight for the Future, ‘Ban Facial Recognition’, (launched July 2019, last accessed September 19, 2023). Link

includes prohibitions on predictive policing AI tools as well as biometric identification and characterisation.[58]

Given the global concern about AI, it is not a surprise that the British government is developing AI policy, and seeking to convene countries that are similarly concerned. It is important that the UK get AI policy right, not only because of the risks, but also because of the opportunities. The UK is one of the world's global leaders in AI research, a standing that could be put at risk by poorly considered regulation.

> **‘The EU's draft AI Act includes prohibitions on predictive policing AI tools as well as biometric identification and characterisation’**

So what is the state of AI policy in the UK, as this global summit approaches?

Well, over the last few years, the Government has released a number of documents outlining approaches to AI regulation. In 2021, it published a national AI strategy in 2021.[59] This strategy emphasised a sector-led approach to regulation that avoided focusing on specific uses of AI. It acknowledged the risks associated with centralised regulation, while also noting that AI would raise difficult issues regarding how regulators handle liability and fairness. The strategy also mentioned the risks posed by AGI and the need for AGI alignment.

This was followed, in 2023, by a White Paper – commissioned due to the surge in public interest, but drawing on the same approach seen in the national strategy.[60] The paper outlined five principles that should underpin a regulatory approach to AI. These were: 1) safety, security and robustness, 2) appropriate transparency and explainability, 3) fairness, 4) accountability and governance, and 5) contestability and redress.

The paper was fairly high-level, with few specific examples. But it contained many encouraging elements. Rather than propose a new AI regulator, the White Paper argued that it would be better to empower existing regulators to develop context-specific approaches to AI. This context-specific approach to regulation is welcome. As argued above, given the wide range of AI applications, it is not appropriate to regulate specific kinds of AI tools or applications. Decentralising AI regulation among dozens of regulatory agencies allows for more industry-specific and context-specific specialisation. It is appropriate for the Medicines and Healthcare products Regulatory Agency to craft guidelines about AI aiding medical diagnoses, and for the Civil Aviation Authority to do the same for AI-fueled navigation applications for drones.

A centralised approach to AI regulatory risks being blunt and resistant to change. It would also likely struggle to account for the nuances of particular AI applications and the fast-paced nature of AI development. Indeed, the Government has thus far been wise to adopt an approach that embraces incremental change. The former head of regulation at the government's Office for Artificial Intelligence has noted that many of the regulations and proposals that the body considered back in 2021 would have

[58] James Vincent, ‘EU draft legislation will ban AI for mass biometric surveillance and predictive policing’, *The Verge* (May 11, 2023). Link

[59] Department for Digital, Culture, Media & Sport, ‘National AI Strategy’, September 2021. Link

[60] Department for Science, Innovation and Technology, ‘A pro-innovation approach to AI regulation’ (March 2023). Link

rapidly become 'ineffective or obsolete' given how the technology turned out, and that 'the UK was wise to adopt an incremental approach'.[61]

The White Paper argued that the aim of the AI regime should be to achieve a number of goals including increasing growth and innovation, instilling public trust in AI, and strengthening the UK's position as a global AI leader. Yet whether that approach will be a success depends on how a wide range of regulators implement new regulations. And the precedents in terms of technology regulation are not encouraging – as I have argued elsewhere, including in multiple papers for the CPS, measures such as the Online Safety Bill or Digital Markets, Competition and Consumers Bill threaten real damage to Britain's competitiveness. Already, the worries about AI noted above are prompting calls for all kinds of regulations and restrictions. No doubt regulators will feel pressure to react from the public and the private sector. And it is striking that the White Paper's proposals task regulators, not DSIT or another department, to craft guidance and advice.

'Decentralising AI regulation among dozens of regulatory agencies allows for more industry-specific and context-specific specialisation'

Similarly, while the decentralised approach embraced by the White Paper is admirable, it is not without risks. Many businesses make products covered by a range of regulatory agencies. It is not inconceivable that companies such as Google, DeepMind, Meta, and Amazon will find themselves governed by a number of conflicting regulators and policies. Such market incumbents have the legal resources to comply with what could be a range of different and mutable AI policies. But smaller firms may struggle if the policies issued by the different regulatory agencies vary too considerably.

In addition, smaller firms will need reassurance that they are not at excessive risk of legal liability. As the Information Commissioner's Office noted in response to the White Paper: 'Businesses will require confidence that implementing any guidance or advice will minimise the risk of legal or enforcement action by regulators. This need is particularly acute for small to medium sized enterprises (SMEs) that may lack the in-house legal expertise of larger organisations. We would welcome clarification on the respective roles of government and regulators in issuing guidance and advice as a result of the proposals in the AI White Paper.'[62]

The White Paper states that the five principles will not be backed up by statute. However, it goes on to say that the government could introduce legislation to impose a duty on regulators to have 'due regard' for the guiding principles.[63] Such legislation could provide clarity that businesses and regulators would welcome – but it would need to avoid becoming a 'Christmas tree' Bill in the style of the Online Safety Bill, stuffed with multiple far-reaching measures in pursuit of often contradictory ends.

[61] Krier, S. [@sebkrier]. (2023, August 14). 'When I reflect on my time at the UK Government's Office for AI, I remember we were considering quite a few regulatory and policy proposals. In hindsight, many of them would have been ineffective or obsolete. So I think the UK was wise to adopt an incremental approach. This shouldn't mean paralysis, but it serves as a reminder to be skeptical of overly confident opinions on what governments should do.'Twitter. Link

[62] Information Commissioner's Office, 'The Information Commissioner's Response to the Government's AI White Paper' (April 3, 2023). Link

[63] Department for Science Innovation and Technology, 'A pro-innovation approach to AI regulation'

There is also the question of how the framework outlined in the White Paper will be monitored from the centre. It may be that a body inside DSIT would be most appropriate. But given that much of the success of the White Paper's recommendations will rest on the central monitoring of regulators, it is crucial for the Government to organise such monitoring in a way that does not sow confusion or regulatory overreach.

A Blueprint for Bletchley

Of course, the White Paper is not the end of the Government's ambitions – and the fact that it is moving ahead with global discussions about AI regulation and AI safety is a welcome step. To that end, in the final section of this paper, I will look ahead to the Bletchley Park summit and outline what a pro-innovation AI safety policy could actually look like.

When considering AI policy, the Government should continue to embrace a framework that is context-specific. As has been noted above, AI is already a ubiquitous feature of modern life and some of its most popular applications, such as foundation models, have a wide range of uses.

> **'Despite the White Paper's welcome contribution to the AI regulation debate, there are still further steps the Government could take'**

It is appropriate for the Government to avoid an approach to AI regulation that seeks to target specific categories of technology powered by AI. Yet despite the White Paper's welcome contribution to the AI regulation debate, there are still steps ministers could take to ensure that regulation does not stifle innovation and that it provides clarity for entrepreneurs, researchers, and investors. These include ensuring that regulators are only tasked with preventing likely and serious harms, while also establishing prediction markets that would allow for government officials as well as the public to gauge the danger of new AI tools and the effectiveness of safety regulation.

Safety charters

That the Government included 'safety, security and robustness' as one of its five principles of AI regulation is not a surprise given the concerns about AI and the widespread attention they receive. However, there is a risk that regulators could be overly cautious when considering AI safety and stifle innovation by recommending burdensome regulations.

The White Paper correctly notes that the AI safety risks include 'physical damage to humans and property, as well as damage to mental health' and states that 'it will be important for all regulators to assess the likelihood that AI could pose a risk to safety in their sector or domain, and take a proportionate approach to managing it.'[64]

One way for the Government to ensure that regulators adopt such a proportionate approach without stifling innovation would be to require regulators to specifically define the harms they are seeking to prevent and the likelihood of such harms occurring. The Government should therefore mandate that each regulator publish an AI policy charter including this information, to be updated as the technology evolves.

[64] Ibid.

Such charters should clearly establish a set of safety standards imposed on regulated AI products and research practices, as well as an outline of the likely and significant harms that they are intended to prevent.

In particular, the AI policy charters should note what significant harms the regulatory agency aims to prevent via its guidance. Death and serious injury are the most obvious, but many regulatory agencies will have to consider significant harms specifically related to the products and practices within their remit. But this should be limited to harms that are genuinely 'significant': those that would result in serious injury or death, the compromising of national security, facilitation of serious crimes, or permanent serious damage to the environment or critical infrastructure.

> ‘The Online Safety Bill is the perfect example of how legitimate concerns can give rise to policy proposals that ignore context and nuance’

For example, although the Government's AI White Paper mentions the potential for AI to affect mental health and spread disinformation, we should be wary of tasking regulators considering harms that stray into the terrain of the subjective, or politically contentious. The Online Safety Bill is the perfect example of how concerns about content moderation and mental health have given rise to policy proposals that ignore the context and nuance associated with these issues. As a result, a regulator - Ofcom - is on the brink of having to oversee dynamic and complex collections of online content, in a way that heavily incentivises firms to limit users' access to legal and valuable content. As I have pointed out in previous CPS papers, the threat of losing 10% of their global revenue will surely incline any tech firm to err on the side of caution, and excessive censorship.[65]

Of course, AI tools and products may well have negative effects on online speech and mental health. But it is not clear that such worries warrant an AI-specific response. In the past, the spread of electricity, the car, the camera, the radio, the telephone and the bicycle all prompted complaints and calls for regulation. No new technology is without costs. But political and social institutions have proven to be robust in the face of them.

Furthermore, given the potential consequences of AI – both good and ill – it is surely right for the state and its regulators to focus their resources on getting to grips with the most serious potential harms, rather than spreading themselves too thinly in trying to regulate the whole digital world into a state of perfect grace.

Therefore, after a regulator has listed the significant harms it seeks to mitigate, it should explain how likely a significant harm is to occur in order to justify intervention. The use of the term 'likely' is of course one that invites subjective judgement. Risk tolerances vary. Some people would rather drive a car from London to Edinburgh than take an aeroplane. Some would happily pay £100 to go skydiving, while others could not be paid £1,000 to do so.

Regulators are in the unenviable position of judging what tolerance of risk is appropriate for the general public. Yet despite the inherent difficulties included in measuring tolerance for risk, there is precedent for regulatory agencies stating what an intolerable risk is. For example, the CAA explained in 2021 that those applying for

[65] Matthew Feeney, 'A Censor's Charter? The case against the Online Safety Bill', *Centre for Policy Studies (September 28, 2022).* Link

launch operator licences, return operator licences, or spaceport licences would have to demonstrate that they had taken steps to ensure that the risks associated with their planned activities were 'as low as reasonably practicable' ('ALARP').[66] The CAA noted that an unacceptable level of societal risk was a risk higher than 1x10-4 of one or more casualties/fatalities per mission.[67]

ALARP is not a new or rare heuristic – it was first codified in the Health and Safety at Work etc. Act 1974. Since then, regulatory agencies and the private sector industry groups and businesses have used it for safety guidance and risk management.[68] It could therefore play an important role in an AI regulatory framework.

> 'Prediction markets force commentators and experts to put moneywhere their mouths are'

Because regulatory agencies cover almost every feature of British commercial life, it would be inappropriate for each to adopt the same central calculation of a tolerable risk of likely harm. Nonetheless, each agency can provide calculations explaining what they consider the likelihood of a significant harm that would justify intervention.

Prediction markets

When household-name technology investors and globally respected researchers sign AI safety commitments or issue warnings of catastrophic AI risk, they make headlines. Unfortunately, as I noted above, these statements are often vague.

One way for the government to establish the UK as a leading AI safety hub would be to build prediction markets. These would enable the public to gauge the threats of new AI tools and how likely the Government – and the experts – consider specific AI harms to be. They would also allow observers to track which AI commentators are more prone to hyperbolic rhetoric than accurate forecasting.

Prediction markets are markets where participants place bets on upcoming events. The prices indicate the likelihood of an event occurring. For example, the New Zealand-based PredictIt allows users to make bets on elections and other political events. Shares of 'Yes' and 'No' trade between $0.01 and $0.99, with the actual outcome paying out one dollar – conveniently meaning that the price reflects the market's probability judgement for a given outcome.

For example, in a US Congressional race a PredictIt user, John Diviner, may believe that the underdog Candidate B is likely to win and sees that on the Predictit market for that race, 'Yes' shares for Candidate B are trading for $0.15 each. Diviner buys 250 shares of 'Yes' for $37.50. Later, it turns out that Diviner was right and Candidate B did win. Predictit redeems Candidate B's shares at one dollar each, resulting in Diviner walking away with a nice profit.[69]

Prediction markets force participants to put money where their mouths are.

[66] Civil Aviation Authority, 'ALARP Acceptability Policy' (July 29, 2021). Link

[67] Ibid.

[68] Nuclear Industry Good Practice Guide, 'The Application of ALARP to Radiological Risk: A Nuclear Industry Good Practice Guide' (2012). Link, Centers for Disease Control and Prevention, 'ALARA – As Low As Reasonably Achievable'. Link

[69] Most of this paragraph is listed from an article on prediction markets I wrote for Works in Progress. Matthew Feeney, 'Markets in Fact-Checking', *WorksinProgress* (February 21, 2023). Link

Economist Alex Tabbarok has remarked that ‘a bet is a tax on bullshit’.[70] The cost for making a bad prediction without a stake is negligible. Many of us are familiar with people in our social and professional lives confidently and boldly making bad predictions. At a time when conversations and debates about AI are often derailed by concerns based on science fiction hypotheticals, it is important that the government seek to elevate the debate. Prediction markets are one way to do that.

> **‘ If the UK government were to establish a venue for AI risk prediction markets, it would send a strong signal to the world that it is serious about AI risk ’**

Prediction markets would also provide useful information to observers who are unsure about how to interpret news about AI safety. Most people who read the news are not qualified to judge the likelihood of a particular AI tool or application causing harm. However, if the Government were to run a prediction market for AI safety, then even the layperson would be able to gauge the likelihood of a new AI tool or application inflicting harm. And regulators would be forewarned about the areas of gravest concern.

For example, a user of the AI safety prediction market could establish a market on whether a particular driverless car would pass a safety test, or whether a new deepfake detection tool will achieve a particular detection rate. Another might choose to bet on whether a new Large Language Model tool would pass the Maths A-Level exam, or whether a facial recognition application will yield the same false positive/negative rates across every racial group.

Of course, there would be many takers for markets asking whether AI would gain sentience, or destroy the world. But it is in the more granular markets that the real value could be found – even if, or even especially if, it turned out that the participants systematically over- or undervalued AI’s progress. Indeed, with an AI risk prediction market in place the public would be better informed about who in government, media, academic, and industry was making the best predictions about AI risk.

If the UK government were to establish a venue for AI risk prediction markets, it would send a strong signal to the world that it is serious about AI risk. It would not be able to help those who want to bet on the likelihood of the most catastrophic events. After all, dead people can’t collect payments. But such markets could nonetheless help inform important debates on a range of AI applications.

Of course, the Government would have to address a number of concerns in order for public AI risk markets to gain widespread acceptance. Some may worry that such markets would allow for insider trading, allowing those working inside firms or the Government to skew the market. One response would be for the Government to require that participating in the AI prediction markets was conditional on submitting identification documents. A user could still use a pseudonym on the market, but the admins would know their identity, thereby reducing the incentive to engage in insider trading.

[70] Alex Tabarrok, ‘A Bet is a Tax on Bullshit’, *Marginal Revolution* (November 2, 2012). Link

Conclusion

Concerns about AI have become a prominent feature of public policy debate and journalism reporting. These range from the relatively minor worries to fears of outright human extinction.

> ‘AI has rich potential to make our country, and the world, a more productive and even a more pleasant place’

It is right to be concerned about these issues, and good that Britain is taking a lead in bringing together leaders, tech firms and thinkers to discuss these issues. But we should not get carried away by the scaremongering. AI has rich potential to make our country, and the world, a more productive and even a more pleasant place. We need to remember to balance the risks with the opportunities – and to adopt a regulatory regime that is flexible, proportionate and treats each case and each sector on its merits.

Centre
for Policy
Studies

57 Tufton Street, London, SW1P 3QL
October 2023
ISBN 978-1-914008-43-6