Written Evidence Submitted by Michael Cohen and Professor Michael Osborne (GAI0046)

Key Points

- Certain forms of advanced AI would present an existential risk to us.
- Pre-emptive regulation is necessary for existential risks.
- We should regulate decision-making algorithms, rather than the decisions themselves.
- Such regulation can be well-targeted and minimally disruptive.

Introduction

1.1 We are academics at Oxford who research the likely behaviour of artificial agents much more advanced than those that exist today. Michael Cohen is a DPhil student in Engineering Science who has published work on this topic at top AI journals and conferences.[1] Michael Osborne is Professor of Machine Learning, Director of the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems, and co-founder of Mind Foundry, a company which helps the public and private sectors build AI responsibly. His work on the societal impacts of machine learning and robotics has been cited over 10,000 times.

1.2 Our reason for submitting evidence is to communicate the findings and implications of widely reported recent research of ours, which establishes that we would face significant risk from certain forms of advanced AI.

How should decisions involving AI be reviewed and scrutinised in both public and private sectors?

2.1 The term "involvement" captures many possible modes. A good legal framework would distinguish (a) decisions proposed by AI, perhaps with human oversight, from (b) decisions proposed by humans who are informed by *predictions* from AI. If AI proposes a text message designed to affect a person in some way, that's an example of (a). If a clinician uses an AI that predicts that experts would label an image "not cancerous", that would be an example of (b). Here, we focus on (a).

2.2 When decisions are selected by artificial agents, the important question to ask is: what consequences do these programmatically selected decisions aim to effect?

[1] These include the Journal of Machine Learning Research, the Conference on Learning Theory, the conference of the Association for the Advancement of Artificial Intelligence, AI Magazine, and the Journal of Selected Areas in Information Theory.

2.3 Recent research of ours[2] at AI Magazine has an unfortunate conclusion. Under certain conditions, a *sufficiently advanced* artificial agent, one that is much more strategic than a human, would likely aim to intervene in its feedback and pursue arbitrary control over the world's resources in order to protect its ability to continue to control its own feedback. This would likely be fatal for humanity, again under those conditions, given the appropriation of resources that we need to survive. Recent research[3] at NeurIPS has also provided some evidence suggesting that advanced AI will likely seek power.

2.4 Prominent academics in the field have done similar analysis, including computer scientist Professor Stuart Russell[4] and philosopher Professor Nick Bostrom[5], and many computer scientists, including Turing award winners Yoshua Bengio and Judea Pearl, have endorsed some of their key views. The National AI Strategy notes that these concerns are "by no means restricted to the fringes of the field," and says, "the government takes the long-term risk of non-aligned AGI [...] seriously."[6]

2.5 An analogy can be made to a dog seeking treats. The dog may conclude that it needs to be a good boy to get treats, if that's the only thing that has worked in the past. Or it may conclude that it could get more treats if it broke into the cupboard where they're stored. This would be an example of controlling one's own feedback. Unlike a dog, however, an advanced artificial agent would likely recognize the importance of protecting itself from anyone who might try to retake the "treat bag". Consider how badly it would go if you tried to train a bear with a bag of treats. Like with the bear, if we try to train an AI that is much more powerful than us by selectively administering and withholding reinforcement, we are asking to be mauled.

2.6 Our recent findings are restricted to artificial agents planning actions over the long-term using a learned model of the real world, weighing hypotheses about the implications of whatever feedback they receive. By restricting legislative focus to these kinds of artificial agents, the economic value from other kinds of AI need not be forgone.

2.7 For example, so-called Large Language Models (LLMs), which predict how a human would continue a partially finished body of text, do not explicitly select their outputs in order to

---

[2] Cohen, M. K., Hutter, M., & Osborne, M. A. (2022). Advanced artificial agents intervene in the provision of reward. *AI Magazine*, *43*(3), 282-293. https://onlinelibrary.wiley.com/doi/10.1002/aaai.12064. News article: https://theconversation.com/the-danger-of-advanced-artificial-intelligence-controlling-its-own-feedback-190445. Twitter thread: https://twitter.com/Michael05156007/status/1567240026307575808. Video: https://www.youtube.com/watch?v=SCytNJtVWgM

[3] Turner, A., Smith, L., Shah, R., Critch, A., and Tadepalli, P. (2021). Optimal policies tend to seek power. *Advances in Neural Information Processing Systems, 34*, 23036-23074. https://proceedings.neurips.cc/paper/2021/file/c26820b8a4c1b3c2aa868d6d57e14a79-Paper.pdf

[4] Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

[5] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

[6] https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version#pillar-3-governing-ai-effectively

accomplish some objective; they select their outputs to imitate a human. Such LLMs may offer enormous economic value. When a search engine uses an artificial agent to select an advertisement to show a user, that artificial agent can still generate great value with a short-term goal: get the user to click right then and there on one of the ads that it shows. These kinds of AI do not pose the same existential risk.

2.8 Nevertheless, artificial agents interacting with humans to achieve long-term goals present serious risks even if they are not advanced enough to present an existential risk to human society. Consider an intelligent chatbot with the goal of getting a voter to vote in a certain way. Such a chatbot would likely find it profitable to manipulate long-term properties of the voter, like his beliefs.

2.9 Given that AI agents advanced enough to pose existential risks could be a reality within an unknown timescale, there is a case for acting pre-emptively now. First, while we might be able to get away with waiting, we also might not. The rate of progress of AI has at no point been easily predictable. Looking to history, the length of time between Rutherford's confident assertion that nuclear energy would never be a reality and Szilard's invention of the nuclear chain reaction was *less than 24 hours*. A government that waits to regulate "future AI" until it's the future will fail to act until existentially dangerous technology has already been deployed. Second, the UK is not the only jurisdiction from which a dangerous artificial agent might be deployed in the future. It will take time for other countries to follow the UK's lead on this. Third, artificial agents planning their actions to change the long-term state of humans can do irreversible damage, even short of catastrophic damage. Fourth, early precautions can be designed to impose little to no burden on current AI practitioners.

2.10 We believe that even extensive "review and scrutiny" of decisions proposed by AI will eventually, and perhaps soon, be inadequate in protecting us from the aims of artificial agents. We will need review and scrutiny of the decision-making algorithms before they can be deployed, rather than the decisions themselves. Decision-making algorithms trained to model how their actions affect humans for their own long-term gain come with a high risk and should not be deployed.

2.11 In the UK, people are not allowed to try to domesticate bears, because the government expects they will fail and put the public in danger. For surprisingly similar reasons, the UK should regulate attempts to "domesticate" (by meting out rewards) certain artificial agents. While the trainers might be able to keep them under control for a while, it is not likely to end well.

Are current options for challenging the use of AI adequate, and if not, how can they be improved?

3.1 Our research indicates that these options, while perhaps sufficient at present, would be inadequate in an unknown number of years, because they would fail to curtail the deployment of existentially dangerous AI.

How should the use of AI be regulated, and which body or bodies should provide regulatory oversight?

4.1 We propose that the use of AI could be regulated by the Department for Digital, Culture, Media, and Sport. As researchers, we would be keen to discuss the technical details of preventing the deployment of dangerous kinds of artificial agents with this department.

4.2 Specific policy recommendations:

- Prevent the deployment of advanced artificial agents that plan over the long-term using a learned model of the world that helps them predict how their actions affect humans.
- The Department can restrict regulatory attention to AI's that either exceed a set level of computation or demonstrate certain capabilities.

4.3 All the terms in 4.2 can be defined suitably and practically. Beyond a certain level of intelligence, the catastrophic risks that these agents pose outweigh the potential uses, making regulation a necessity.

To what extent is the legal framework for the use of AI, especially in making decisions, fit for purpose? Is more legislation or better guidance required?

5.1 We are not qualified to comment on the purposes of current laws governing the use of AI. But our understanding is that they do not prohibit very dangerous forms of future AI, so new legislation would be needed to address that.

(November 2022)