Purpose of the Anti-Artificial Scheming Act

Michael Cohen
DPhil Candidate, Engineering Science, University of Oxford
January 2023

1. Risks from novel advanced AI

1.1 The existential risk arising from novel advanced AI has been outlined and explained in a submission of evidence to the House of Commons Science and Technology Select Committee inquiry on the Governance of Artificial Intelligence[1] based on the findings of Cohen, et al. (2022) and others.[2] This document presents a more detailed proposal of what to do in light of those findings, and it is aimed at policymakers who have read that submission of evidence.

2. Explanation of the Anti-Artificial Scheming Act

2.1 The Anti-Artificial Scheming Act aims to prevent the training and deployment of artificial agents that a) are very advanced, b) act to achieve long-term goals, and c) face an incentive to intervene in the protocols designed to control them. And the Act aims to put minimal regulatory burden on other forms of AI.

2.2 Pursuant to the Anti-Artificial Scheming Act, certain AI projects would require a license from the Department of Digital, Culture, Media, and Sport (DCMS).

2.3 In order to put minimal burden on artificial agents that are not very advanced, licenses would only be required for artificial agents whose model of the world was extensively trained with significant computing resources.

2.4 In order to put minimal burden on artificial agents that only have very short-term goals, licenses would only be required for artificial agents that pick their actions in pursuit of a long-term goal. For example, the program that decides what links to display after a Google search could be called an artificial agent that pursues the short-term goal of getting one of the links clicked. Importantly, the existentially concerning results of Cohen, et al. (2022) do not apply to agents whose only goals are immediate.

2.5 The Anti-Artificial Scheming Act directs the DCMS to only grant licenses to certain advanced artificial agents that act to achieve long-term goals. The Act instructs the DCMS how to craft regulation such that licenses are denied to agents that are likely to face an incentive to intervene in the protocols designed to control them, without requiring the DCMS to evaluate this question directly.

2.6 The results of Cohen, et al. (2022) only apply to artificial agents that use machine learning to understand how the world responds to their actions, and so the Act directs the DCMS to grant licenses to artificial agents that do not use machine learning. For example, Google search originally did not use machine learning when deciding what links to display.

[1] https://committees.parliament.uk/writtenevidence/113797/pdf/
[2] https://onlinelibrary.wiley.com/doi/10.1002/aaai.12064

2.7 Finally, Assumption 3 of Cohen, et al. (2022) likely fails if the artificial agent in question does not understand how its actions affect humans, so these agents do not present the same existential risk to us. Therefore, the Act directs the DCMS to grant licenses to artificial agents that cannot model how they might use humans as means to an end. The Act gives more detailed guidance in how to construct such a rule in a practical way.

3. Costs of the Anti-Artificial Scheming Act

3.1 The DCMS must direct resources to granting licenses, and companies must direct resources to applying for them.

3.2 Even when correctly following the provisions of the Act, the DCMS may deny licenses to artificial agents that would have been harmless and economically useful.

3.3 However, significant resources seem to be required to train agents that are very advanced. Since small projects can proceed with licenses, the costs represented in 3.1 likely represent a small fraction of the costs of creating the AI in the first place.

3.4 As of early 2023, the costs described in 3.2 are, to our knowledge, hypothetical. We are not aware of any current industrial uses of AI which would fail to receive a license from the DCMS according to the Anti-Artificial Scheming Act, let alone any economically critical uses of AI. This is not to say that the costs described in 3.2 are nil, only that they do not include any upheaval of British industry.

4. Costs of not implementing the Anti-Artificial Scheming Act

4.1 On our current trajectory, once artificial agents become sufficiently advanced, if they are designed in the wrong way, they will present of a large risk of causing human extinction. If the major governments of the world do not implement the Anti-Artificial Scheming Act or something like it, and there is no other Plan B, this is the cost.

4.2 We have not heard of any alternative regulatory proposals, let alone compelling ones.