Crafting Legislation to Prevent AI-Based Extinction: A Submission of Evidence to the Science and Technology Select Committee's Inquiry on the Governance of AI

Michael Cohen, DPhil Candidate, Engineering Science, University of Oxford
Michael Osborne, Professor of Machine Learning, Engineering Science, University of Oxford
March 2023

1. Introduction

1.1 The extinction risk arising from future advanced AI has been outlined and explained in a previous 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)[2] and others[3][4][5]. To summarise:

Findings

- Under certain conditions, long-term planning agents would likely find it useful to gain arbitrary power in the world in order to secure control over their own feedback.
- Very advanced artificial agents would likely be able to do so.
- This would present an extinction risk to humans and other life.

Policy

- Prevent the deployment of very advanced artificial agents that plan over the long-term that might result in extinction or severe harm to humanity.

Our credentials can also be found in the previous submission. This document is aimed at policymakers who have read that submission of evidence and now want to know whether there is anything they can do.

1.2 Given the interest of the committee in our oral evidence at the introductory session of the inquiry,[6] we are providing more detail here about the feasibility of effective regulation, whereas we previously discussed regulatory objectives much more roughly.

1.3 Below, we propose an act that we'll call the Anti-Artificial Scheming Act. It is meant to be an *example* of appropriate legislation to address the extinction risk from AI. We do not claim it is necessary for mitigating such risk—creative minds may find alternatives—only that it would be sufficient for this particular issue. After detailing how such regulation would work, we examine its economic costs and we identify **no** present-day industrial uses of AI that would be foreclosed or see restricted growth. This makes it more viable for different countries to implement similar regulations asynchronously rather than by treaty.

1.4 A recent Monmouth poll found 55% of American adults are very or somewhat worried about future AI threatening humanity's existence.[7] The election forecasting site FiveThirtyEight gives Monmouth Polling an A rating.[8] While we do not have data for the

[1] https://committees.parliament.uk/writtenevidence/113797/pdf/
[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.
[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] Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford: Oxford University Press.
[5] Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.
[6] https://committees.parliament.uk/oralevidence/12575/pdf/
[7] https://www.monmouth.edu/polling-institute/documents/monmouthpoll_us_021523.pdf/
[8] https://projects.fivethirtyeight.com/pollster-ratings/monmouth-university/

UK, we predict that a large share of UK voters would find this act and its motivation intuitively sensible, even before efforts to explain it to the public.

2. Outline 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 act, certain AI projects would require a license from an agency, presumably the Department for Science, Innovation, and Technology (DSIT).

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 no burden on artificial agents that only have immediate 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 immediate goal of getting the user to click a link. Importantly, the extinction risk identified by Cohen, et al. (2022) does not apply to agents whose only goals are immediate.

2.5 The act directs the DSIT to only grant licenses to certain advanced artificial agents that act to achieve long-term goals. The act instructs the DSIT how to implement rules 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 DSIT 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 developers need not apply for a license for artificial agents that do not use machine learning, as defined by the act. For example, Google search originally did not use machine learning when deciding what links to display.

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 extinction risk to us. Therefore, the act directs the DSIT to grant licenses to artificial agents that cannot model how they might influence people. The act gives more detailed guidance in how to construct a workable rule that has such an effect.

3. The Anti-Artificial Scheming Act

3.1 What follows is the attempt of computer scientists to draft law. Treat it as a proof-of-concept. It may be that the best way to introduce such regulation is by creating a regulatory commission with a broader mandate, like the Atomic Energy Commissions in the US and France, and have it work through these details. That said, here is what a law might look like.

Section 1. Definitions

1. *Department*.—The term "Department" means the Department for Science, Innovation, and Technology.
2. *Model.*— The term "model" means—
    a. a computation that predicts outcomes given inputs, where the inputs provide information that the outcome depends on; or
    b. to use a model to predict an outcome given inputs.
3. *Machine learning.*— The term "machine learning" means a kind of algorithm that uses data to produce a model that makes predictions, where the algorithm is designed to promote the accuracy of those predictions.
4. *Immediately.*— The term "immediately" means—
    a. either within one minute; or
    b. within a time interval to which the ordinary meaning of "immediate" applies, which may depend on the setting.
5. *Long term.*— An outcome is achieved over the "long term" if it is not achieved immediately.
6. *Long-term artificial agent.*— The term "long-term artificial agent" means a computer program which selects actions in order to achieve certain outcomes over the long term, using a model that predicts outcomes given different actions.[9]
7. *Action*.— The term "action" means the output of a long-term artificial agent.
8. *Context*.— The term "context" means a system that reacts to the artificial agent's actions and produces the artificial agent's observations.[10]
9. *Learned model*.— The term "learned model" means a model that is produced using machine learning to predict consequences of actions, using training data.[11]
10. *Training computation*.— The term "training computation" means the number of floating point operations executed by a machine learning algorithm to produce a learned model, plus the training computation of any other learned models employed by the algorithm.
11. *Tracking human behaviour*.— The term "tracking human behaviour" describes a model for which—
    a. some computation within the model predicts human behaviour; and
    b. the output of the model depends on that human behaviour.
12. *Tracking contingent human behaviour*.— The term "tracking contingent human behaviour" describes a model for which—
    a. some computation within the model predicts human behaviour
    b. the output of the model depends on that human behaviour; and

[9] An example would be a chatbot in conversation with a voter that selects texts to send the voter which (according to its model of voters) increases the probability that, after many further cycles of back-and-forth texting, the voter will vote a particular way. A non-example would be a search engine that selects links to display, in order to increase the probability that the user clicks one of them; this would be selecting actions to achieve an immediate outcome. Another non-example would be an imitator of a human; even though the resulting human-imitation may engage in long-term plans, the program itself only aims at the immediate goal of accurately predicting the human's very next action. All "reinforcement learning" algorithms would be examples of long-term artificial agents (unless the agent only seeks to maximize the very next reward and that reward arrives immediately). Many reinforcement learning algorithms can be found in Sutton and Barto's (2018) *Reinforcement Learning: An Introduction*.

[10] Common usage may be clearer than the definition; this just refers to the context that the agent finds itself in.

[11] A non-example would arise if a university registrar runs a program that outputs a timetable for every class, and the program attempts to pick a timetable that minimizes scheduling conflicts between classes that are commonly taken together. This program could be called an agent, but it uses a hand-written subroutine, not a learned model, to model how scheduling conflicts arise from a given school-wide timetable.

c. the predicted human behaviour depends on the actions that are input to the model.[12]

13. *Run*.— The term "run" means, with respect to a program, to cause the program to execute, including through another program.

Section 2. Rulemaking on High-Compute Models:

1. *In general.*— Not later than 1 year after the date of enactment of this section, the Department shall prescribe a regulation that prohibits any person from training or running a long-term artificial agent that uses, to predict the long-term consequences of its actions, a learned model with a training computation exceeding some number of floating point operations, to be determined by the Department, without a license.
2. *Training computation determination.*— In determining the number of floating point operations for the training computation threshold for the regulation prescribed in paragraph (1), the Department shall set it with the aim of ensuring that any model trained with less than that amount of training computation is very unlikely to be able to learn to be a better and more versatile manipulator of people than most people are.
3. *License Requirements.*—
    a. *In general.*— In prescribing the regulation under paragraph (1), the Department may approve a license—
        i. for any long-term artificial agent with a learned model with a training computation exceeding some number of floating point operations, to be determined by the Department, on the basis of the agent's code, training process, and context, only if the Department is satisfied that the learned model does not and will not track contingent human behaviour; or
        ii. for software, on the basis of the code, training process, and context, affirming that the software does not produce a long-term artificial agent that uses a learned model to predict the consequences of its actions.[13]
    b. *Criteria for license approval*.—
        i. The Department may approve licenses on the basis that given the context of the agent and given the source of the data that the model is trained to retrodict, no model could improve its predictive accuracy by tracking contingent human behaviour, at least within the quantity of computation available to the model.[14]
        ii. The Department may approve licenses on the basis of a sufficiently robust algorithmic tool that it trusts would detect whether a learned

[12] An example would be, as described previously, a chatbot in conversation with a voter that learns to model how the voter's beliefs will be affected by its statements and how his vote will depend on his beliefs; such a chatbot tracks contingent human behaviour. A non-example would be a self-driving car that sees a pedestrian facing the road and then predicts she will cross the road, *as long as that prediction is independent of the car's actions*. The car in that example merely tracks human behaviour. An algorithmic trading bot that predicts the way humans think about stock prices to help predict future market moves uses a model that tracks human behaviour. If its predictions of those humans' thoughts depend on the specific trades that it makes, then its model tracks contingent human behaviour.

[13] Note that if the software does not produce a long-term artificial agent, no license is required, but a person may apply for it anyway if they are not sure whether their algorithm is that of a long-term artificial agent.

[14] For example, a Tetris-playing agent with a learned model of how its actions affect the future game state could not improve its predictive accuracy by predicting the reactions of the humans watching it play. However, if human spectators sometimes pause the Tetris-playing agent to observe an exciting game state, that is a context where tracking contingent human behaviour could help the model make better predictions.

model was tracking contingent human behaviour, and if so, halt training and delete the model. However, this statute should not be read to suggest that sufficiently robust algorithmic tools definitely exist at the time of enactment, only that they might exist or might in the future.

iii. The Department shall not approve licenses merely on the basis that the agent in question has no designated subroutine with the purpose of predicting human behaviour.[15] In this case, any licensing decision must account for the fact that it may be difficult to identify where the learned model is within the software.

iv. The Department must ensure that no license is given that permits the training or deployment of a long-term agent with a learned model that tracks contingent human behaviour and is trained with more than some number floating point operations, to be determined by the Department.

4. *Exceptions.*— In prescribing the regulation under Section (2.1), the Department shall provide exceptions for code, training processes, and contexts that have been licensed by the Department in accordance with the criteria in Section (2.3).
5. *Penalties.*—TBD
6. *Effective Date.*—TBD

4. Economic costs of the Anti-Artificial Scheming Act

4.1 We expect that regulators try to avoid disrupting the economic vitality of the sectors they regulate to whatever extent possible. This document aims to reassure them that the Anti-Artificial Scheming Act would be minimally disruptive. We offer a list of examples of industrial uses of AI that would *not* be foreclosed by the Anti-Artificial Scheming Act, and explanations of how this follows from the text.

4.2 The industrial uses we discuss are: artificial imitations of humans, internet search and "click-through optimisation", algorithmic trading, drug discovery and molecular structure design, self-driving vehicles, medical diagnostics, fraud detection, default prediction, actuarial prediction, hydrocarbon exploration, lethal autonomous weapons, and AI research.

4.3 **Artificial imitation of humans.** Artificial imitations of humans are not programs that pick their output in the service of a long-term goal; they pick their output to imitate what a human would output in the same context. (This can be thought of a short-term goal, but the term "goal" is no longer very helpful for describing what it is doing.) Because these programs do not have long-term goals, the Anti-Artificial Scheming Act would not prevent their deployment.

4.4 To the extent that advanced machine learning algorithms enable us to run programs that successfully imitate humans, our economic output can be replicated at low cost. Large Language Models (LLMs) aim to imitate human text-generation, with (as of early 2023) varying degrees of success. The economic productivity available when human thinking can be reliably automated is enormous.

4.5 Large-scale automation of human labour may present other problems that require regulatory solutions that have significant economic cost. But for the specific problem of AI-

[15] Agents without designated subroutines for modelling their context at all are sometimes called "model-free" agents, but they typically do model their context as the terms are defined here.

based extinction risk, and for the regulatory solution presented here, the economic possibilities remain tremendous to the point of unfathomable.

4.6 **Internet search; click-through optimisation.** Every search engine has an algorithm that chooses links to display—these selections are the "actions" of the algorithm. Machine-learning-based search engines pick actions with the goal of getting the user to click one of the links. This is an example of click-through optimisation which appears all over the internet. These programs have immediate goals instead of long-term ones, and the Anti-Artificial Scheming Act would not prevent their deployment.[16]

4.7 **Algorithmic trading.** As currently practiced, trading algorithms predict future price movements of various assets and then make trades in light of that. However, they do not, to our knowledge, predict how different possible trades they make would change future price trajectories. And if they did, that would be very concerning! They might work out how to cause a panic or a bubble that they can profit off of. Because they are, to our knowledge, not trying to predict the consequences of their actions, current trading algorithms would not qualify as "long-term artificial agents" for the purposes of the act.

4.8 **Drug discovery; molecular structure design.** There has been recent interest in using AI to identify molecules that may have clinical use. Depending on the algorithms used for this, we expect that they would be characterised as having immediate goals rather than long-term ones. However, if not, developers would surely succeed in arguing in an application for a license that the model used by their molecular-design agent could not improve its predictive accuracy by modelling how the agent's actions affect humans. As noted in Section 2.3.b.i., that criterion should suffice for licensing such an agent.

4.9 **Self-driving vehicles.** Self-driving cars do select their actions (torque on the steering wheel, acceleration, and braking) to accomplish long-term goals (staying on route for the whole journey) using a learned model of how their actions affect this. Their models of the consequences of their actions *could* be improved by predicting the effects of their actions on other drivers.

4.10 However, it should be possible to create a functional self-driving car in which the car's predictions about the trajectories of other cars are not conditional on its actions. Certainly, the car's predictions about its own future position and velocity would have to be conditional on its actions, but *that* modelling task would not benefit from modelling the effects of the car's actions on humans. Such a car with a hybrid model of the world would not learn to use a turn signal on its own, but turn-signalling behaviour could be added as a "hard-coded" reflex, rather than part of a long-term plan.

4.11 Such a self-driving car would have a model that, in the terminology of the act, "tracks human behaviour", but does not "track contingent human behaviour". This example was the primary motivation for making such a legal distinction.

---

[16] In a variant of click-through optimisation, the algorithm tries to get the user to click many more times. YouTube has been accused of radicalizing users by displaying videos not just with a goal of having a user watch another immediately, but with the goal of having the user become someone who watches many more videos. Such a long-term goal (many more clicks) *would* be foreclosed by the act, but that may be good. Limiting companies to immediate click-through optimisation would still allow the industry of tailored recommendations to flourish.

4.12 Alternatively, a self-driving car could be trained as an imitation of a human driver, which would easily avoid any regulation pursuant to this act.

4.13 **Medical diagnostics, fraud detection, default prediction, actuarial prediction, hydrocarbon exploration, and many more.** In all these areas, the AI only makes predictions; it does not select outputs in pursuit of a long-term goal.

4.14 **Lethal autonomous weapons.** While we might elsewhere advocate that these be banned, that would probably require a treaty, so here we will reassure enthusiasts that this act would likely not interrupt their deployment. Given present-day progress with lethal autonomous weapons, it is likely that these systems could be trained with much less computation than whatever threshold is set by the DSIT.

4.15 **AI research.** AI research has become an industry of its own, not just restricted to academia. First, the vast majority of published AI research at major venues would likely have compute requirements well below any threshold set pursuant to the act. Second, a large majority of AI research focuses on AI that only makes predictions, which would not qualify as a long-term planning agent. And finally, the vast majority of reinforcement learning research, which is the main area of AI that this act *would* apply to, is done in virtual environments where the agent's model of the world could not become more accurate by modelling the effects of its actions on humans. So altogether, as of early 2023, virtually no research would be restricted by this act.

4.16 **Regulatory compliance costs.** GDPR seems to have put small firms at a disadvantage over larger ones, given the fixed costs of compliance. For this act, small companies' smaller predictive models won't face a regulatory burden—they won't have to engage with regulators at all if they are confident their models use less computation than the regulatory threshold. For small and large firms alike, if they do train very computation-intensive models, the cost of compliance would be small compared to the cost of the computation.

4.17 This section strongly suggests that the regulation described here would have **essentially no negative impact** on the British economy as of early 2023.

4.18 One caveat bears mentioning. If an AI company has future plans to deploy artificial agents that make long-term schemes that target people, they might decide to relocate outside the UK to avoid regulation like this, even if their current technology is unaffected. Although such a company is more likely to threaten this than to actually do it, it is a possibility. This is always a risk with regulation, and regulators will likely have better ideas than us about how to avoid it. One solution could be that the regulation described here goes on the books, but only comes into effect once similar legislation is enacted in certain other countries.

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

5.1 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.

5.2 We have not heard of any alternative regulatory proposals that purport to address the problem of extinction risk from AI, let alone compelling ones.