Outline of the Anti-Artificial Scheming Act

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

Introductory apology: I am no legislative drafter. This should communicate the thrust, but I suspect that significant rethinking is necessary to create a solid legal framework.

Summary: The act aims to prevent the creation of artificial agents that plan how they can achieve certain outcomes over the long term, while capable of forming a strong understanding of how they can influence people. Developers of high-compute artificial agents must first apply for a license if their agent is designed to plan over the long term. An agency (I propose the DCMS) must determine whether the software in question complies with the last point—not plausibly producing an agent with a strong understanding of how to influence people. The act offers the agency specific and workable guidelines for to determine this. The key provisions are in Sec. 2.1 and 2.3.a.i.

Sec. 1. Definitions:

They are most easily read in the following order: 2, 8, 7, 5, 6, 1, 3, 4, 12, 11, 10, 9.

1. *Action*.— The term "action" means the output of a long-term artificial agent.
2. *Department*.—The term "Department" means the Department of Digital, Culture, Media, and Sport.
3. *Context*.— The term "context" means a system that reacts to the artificial agent's actions and produces the artificial agent's observations.
4. *Learned model*.— The term "learned model" means a model that is produced using machine learning to predict consequences of actions given training data.[1]
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.[2]
7. *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.
8. *Model.*— The term "model" means—

[1] For example, a chatbot in conversation with a voter learns to model how conversations change voters' votes, given past examples of conversations doing that successfully. A non-example would be a university registrar running 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. A simple program, not a learned model, models how scheduling conflicts arise from a given school-wide timetable.

[2] For example, a chatbot in conversation with a voter that selects text messages to show 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 a short-term outcome.

    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.
9. *Run*.— The term "run" means, with respect to a program, to cause the program to execute, including through another program.
10. *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
    c. the predicted human behaviour depends on the actions that are input to the model.[3]
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. *Training computation*.— The term "training computation" means the number of floating point operations executed to produce a learned model, plus the training computation of any other learned models employed in the training process.

Sec. 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 training computation exceeding some number of floating point operations for purposes of compliance with 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 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, and with due notice revoke, a license—
        i. for any long-term artificial agent with a learned model with a training computation exceeding some number 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

[3] For example, a chatbot in conversation with a voter learns to model how the voter's beliefs will be affected by its statements and how his vote will depend on his beliefs, so it tracks contingent human behavior. A non-example would be a self-driving car that sees a pedestrian facing the road and leaning forward and 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 behavior. 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 behavior. If its predictions of those humans' thoughts depend on the specific trades that it makes, then its model tracks contingent human behavior.

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.

b. *Requirements for license approval*.—
   i. 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.[4]
   ii. 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.[5]
   iii. The Department may approve licenses on the basis of sufficiently robust technical analyses of a long-term agent's model that prove to the Department's satisfaction that it is not possible for the model to be tracking contingent human behaviour. However, this statute should not be read to suggest that sufficiently robust technical analyses definitely exist at the time of enactment, only that they might or might in the future.
   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] Agents without designated subroutines for modeling their context at all are sometimes called "model-free" agents, but they typically do model their context as the terms are defined in this law.

[5] 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 behavior could help the model make better predictions.