Industrial uses of AI not affected by the Anti-Artificial Scheming Act

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

I imagine 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. I offer a list of examples of industrial uses of AI that would *not* be foreclosed by the Anti-Artificial Scheming Act.

1. **Artificial imitations 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.

   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.

   Whether such an outcome is on balance good is a question for another time; if not, some may favour protectionist legislation to prevent people's income from going the way of horses' income post-combustion-engine, although I would instead favour increased redistribution.

   Regardless, the economic possibilities still available from AI with the passage of this Act are tremendous to the point of unfathomable.

2. **Internet search; click-through optimisation**

   Every search engine has an algorithm which 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; despite its bad reputation at times, it can be very useful and economically valuable.

   These programs have immediate goals instead of long-term ones, and the Anti-Artificial Scheming Act would not prevent their deployment.

3. **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 my 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 my 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. **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, I expect that they could 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.ii., that criterion should suffice for licensing such an agent.

5. **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) 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.

   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.

   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".

   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.

6. **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.

7. 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 “supervised learning”—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.