A(G)I Safety & Regulation

Views for and from smaller innovating stakeholders:
University spin-offs, foundations, deep tech start-ups, SMEs

Prepared by

Version 1.0 (11-2023)

AT OXFORD IMMUNE ALGORITHMICS (OIA), we have been thinking deeply of the impact of AI and AGI regulation. OIA applies AI at multiple levels; predictive, generative, hybrid and AGI, and is at the forefront of AI innovation pioneering applications from and to cell and molecular biology. We strongly believe our position is not exclusive to us and other stakeholders play an important role in innovation from scholars to start-ups to industry adopters, deserve to be heard. We have collected some points that we think are relevant to what we think should be strategies to be taken into consideration when discussing AI safety and regulation.

1. Tiered and Staged Regulatory Approach

A tiered approach to AI regulation recognises the varying levels of risk posed by different companies based on their size and user base. For instance, a start-up in the initial stages of development, with no user base, poses minimal public risk. Hence, imposing heavy regulations on such companies could hinder innovation and deter investors. In contrast, larger corporations, with a vast user base, should be under stringent regulation and supervision to ensure public safety. There have been some suggestions in the US along these lines but nothing as clear.

Regulation should be tailored to the specific risks and outcomes of AI applications. The focus should be on the impact and potential harm of the AI's application rather than on blanket regulations covering all AI technologies.

Example 1: A small AI start-up developing a new medical diagnostic tool should not be subjected to the same regulatory scrutiny as a tech giant deploying AI in widespread public applications like facial recognition, nor even a medical device corporation that can afford long and expensive regulatory processes.

Example 2: Regulating an AI system designed to synthesise chemical compounds should be more stringent than regulating a basic AI-driven search engine. The potential for harm in the former is significantly greater. There is some indication that some jurisdictions may adopt this approach, including some state initiatives in the US and the European Union although it is quickly evolving and on occasion this has been backtracked to some extent.

2. Small Player Incentives and Levelling the Playfield

Increase Tax on corporate profits for the use of public data and pay content generators a tiered universal income (the more you have contributed with quality input the more you earn, e.g. newspapers, book writers). Consider increasing R&D tax credits from the additional tax revenue from this source.

Level the field for small benevolent actors and create an environment increasingly hostile for possible large malevolent actors to defraud by increasing supervision, regulation and even fees and taxation on large companies to provide grants to smaller ones.

3. Regulate Its Application, Not R&D

Current and future AI is far from having any benign or malign objectives that would not exist today (and already do harm through other means spearheaded by social media). It seems therefore that people's use is what has to be regulated and laws to be enforced.

The focus is on regulating the use-cases of AI, particularly in sensitive fields like medicine, rather than the underlying technologies like Large Language Models (LLMs).

Example 1. Biden's recent Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, suggests that tracking energy consumption from training large models will be key to detect large models subject to regulation. This will embed placing regulation at the R&D stage where avoidance is significantly easier. For example, a start-up distributing energy consumption across many jurisdictions using a technique called Federated AI that requires training smaller neural networks and then sharing weights (what has learned) to a larger upper neural network effectively having the same result as training a huge LLM in a single place but without detection.

Example 2: In medicine, the emphasis is already on the risk-benefit analysis of AI applications in patient care, rather than imposing limits on the size or energy consumption of AI models. The AI component of any solution should be irrelevant as long as an exhaustive risk-benefit analysis is conducted for every application in different circumstances such as normal, exceptional or extreme use cases.

4. Monitor AI with equally powerful AI

As AI systems become more complex and faster, human-only monitoring becomes inadequate. Utilising AI to monitor AI, particularly through adversarial networks, can ensure more effective and thorough regulation.

Both general and domain-specific test batteries with adversarial LLMs should be made publicly available for new released versions and following updates for each kind of Generative AI technology such as LLMs or image-based, independent of conflict of interests or dominion of large corporations. Probably even led by the small players not the large ones. What is an adversarial LLM? The best is to use AI against itself. Turn AI against each other, Generative adversarial NNs can explore a distribution space of potential malevolent use, cases which have not yet devised by humans themselves.

Example: Employing adversarial neural networks (NNs) to test and identify potential biases or ethical issues in new AI models before they are deployed. These adversarial NNs would be supervised by human regulators to ensure comprehensive oversight.

5. Government-Subsidised Technical Infrastructure

This point advocates for government support in levelling the playing field between smaller, benevolent AI developers and larger, potentially malevolent actors. This could be achieved through increased supervision and regulation of large companies, coupled with subsidies and grants for smaller entities.

Example: Imposing higher regulatory fees on large tech corporations, which could be used to fund grants for smaller AI start-ups focusing on ethical and beneficial AI applications. Many countries have an insufficient research grant system that often creates the wrong incentives promoting power concentration in small groups that only give the impression of progress and are only incremental in their advancements. The grant system has become too risk averse and does not truly incentivise radical innovation and moon shots except in exceptional cases that make the system even more unfair by awarding only a handful of multi-million dollar grants often to groups that had already access to appropriate funding. This is especially worse in Europe and the UK than the US and now even China.

6. Well-defined Definitions & Scope Boundaries

Not all AI is created the same or for the same purpose, and not all AI should be regulated as AGI probably should. While the OECD has proposed a definition of AI that regulators seem to be adopting, there seems to be less consensus on the definition of Artificial General Intelligence. The OECD definition of AI reads:

An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

Example: What is AGI?

In the same spirit, to keep AGI away from anthropomorphising its definition and independent of any performance measure, we propose the following:

An AGI system is a machine-based system that approximates a multi-modal and multi-model universal predictor that may or may not receive any input but, for a given piece of information, if required to take an action to simulate or execute its continuation, it does so by approximating a theoretical optimum.

Under this definition, for example, GPT 4 and other LLMs are clearly not AGI yet not only because they are not truly multimodal but most important because we know they are not optimal predictors by design. To pick the next most likely word according to a statistical distribution is not.

Humans are borderline general intelligence on purpose, they would be at the AGI border if they were seen as machines, according to this definition. Given that they are multi-modal (with still a strong anthropomorphic component in the data type) and estimators capable of general prediction (most likely imperfectly), in line with our mind capacity and inability to instantiate optimal logical inference on a regular basis.

Multi-model means that any AGI system should be capable of producing several models to choose from, and to be able to select the best model according to the information available. AIs that produce only one model or pick the model without observing intrinsic information about, e.g. by statistical choice (as a function of 'temperature' in LLMs), cannot be AGI.

The weight of the definition therefore relies mostly on the theoretical optimal definition of inference and prediction. Under this definition, current statistical AI and Machine Learning, including LLMs and current Generative AI approaches, are not AGI and may never be unless they implement some advances that may or may not be breakthroughs.

Does the way a human mind operate show any sign of being more than a version or instantiation of statistical machine learning? There are strong signals that do suggest they do not operate like statistical machines. Their energy efficiency is nowhere comparable to large energy-hungry deep learning approaches including LLMs. This is based on evidence of an algorithmic theory of cognition based on experimental data (Gauvrit, Zenil, et al. 2017) and how animal cognition are approximations under the same theoretical framework (Zenil, Marshall and Tegner, 2023).

In conclusion, these proposals suggest a balanced and dynamic approach to AI regulation, prioritising innovation while ensuring public safety and the ethical use of well-defined AI technologies. This is an attempt to de-anthropomorphise definitions as the implementation of regulating measures would require careful consideration of technical and conceptual aspects never before faced. This related to the debate about whether technology is human-like or not and humans and inanimate matter fall into overlapping definitions.

References and relevant material to read

- N. Gauvrit, H. Zenil, F. Soler-Toscano, J.-P. Delahaye, P. Brugger, Human Behavioral Complexity Peaks at Age 25, PLoS Comput Biol 13(4): e1005408, 2017.

- H. Zenil, J.A.R. Marshall, J. Tegnér, Approximations of Algorithmic and Structural Complexity Validate Cognitive-behavioural Experimental Results, vol. 16 Frontiers In Computational Neuroscience, 2023.

- FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

- The AI regulations that aren't being talked about, Deloitte, 2023. https://www2.deloitte.com/us/en/insights/industry/public-sector/ai-regulations-around-the-world.html?id=us:2sm:3li:4diUS176826:5awa:6di:MMDDYY:author&pkid=1011893

- Artificial Intelligence Index Report 2023, https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

- UnitedHealth uses AI model with 90% error rate to deny care https://apple.news/Ae2cDA6wQSk2nETrocSL3Ig