Campaign for AI Safety

Response: Encouraging Innovative Technologies, Services, Use Cases, and Business Models through Regulatory Sandbox in Digital Communication Sector

This submission is made by the Campaign for AI Safety in response to the Telecom Regulatory Authority of India, GOI's Encouraging Innovative Technologies, Services, Use Cases, and Business Models through Regulatory Sandbox in Digital Communication Sector consultation paper. We trust this submission is of assistance. The Campaign for AI Safety is a not-for-profit association established in Australia with members across the globe. We are concerned about the dangers AI poses to humanity and advocate for a stop to the advancement of AI capabilities and regulation that prioritises safe and responsible AI. We are not affiliated with any political group. Please visit campaignforaisafety.org for more information.

- We write our submission within the context of artificial intelligence (AI) technology.
- The draft framework does not include demonstrating that a new technology can be safely tested. In our view, safety must be the number one priority for testing cutting-edge AI technology. The 'black box' problem (developers do not know how it work inside) and AI systems' capacity to 'learn' from data and change the way it makes decisions or takes actions makes pose risks such as losing control which can have devastating consequences, and unknowingly breaking laws or infringing on individual rights.
- The discussion paper does not explicitly discuss resourcing or funding the proposed framework. We are concerned that, without adequate resourcing of organisational capacity and capability (e.g. AI or emerging technology expertise), regulators will be exposed to undue influence and reliance on regulatees that have greater technical knowledge in these sandboxes and testbeds.
- We are concerned that close collaboration with industry will influence regulators to give more weight to private preferences at the expense of the broader public interest. For example, the AI industry is dominated by the most powerful firms that can act as quasi-regulators in the markets they operate[1].

In the context of regulating AI, we make three policy recommendations for the Government of India's consideration.

[1] CMA response to DCMS pro-innovation approach for regulating AI, Competition and Markets Authority, United Kingdom page 5

1. Prohibit developing even more powerful AI until it is shown to be safe

Today's "foundation models"[2] are trained using substantial amounts of computational power on vast quantities of data (such as the "Common Crawl"[3] of the Internet). Already existing foundation models contain more "parameters" (artificial neurons) than the number of actual neurons in a human brain[4]. As a result, they can perform a wide range of tasks and applications from writing computer programming code to answering questions using their vast internal knowledge to basic reasoning.

Many leading AI researchers believe[5] that these models can eventually be trained to be far more intelligent than any human, in the same way humans are far more intelligent than any ape. Many warn that the consequences could be as catastrophic for humans as humans have been for apes.

Others warn that AI does not even need to have human-level general intelligence. A sufficiently advanced but narrow AI (for example, one trained to perform cybercrime and targeted at critical infrastructure) can be used by malicious actors to create a catastrophe[6].

At the same time, these models are "black boxes": AI labs know how to train them, but do not currently understand how they work inside[7]. Many capabilities of foundation models are "emergent", and as such cannot be predicted simply by extrapolating the performance of smaller models[8]. Furthermore, these models can have additional capabilities from "programmatic scaffolding"[9] created by third parties.

Thus, in expert opinion, already powerful "black box" AI systems can be made devastatingly and uncontrollably dangerous.

It is a **commonly held view that the only way to ensure safety is to delay the development of AI advancing towards human intelligence** until it is proven to be safe and to research the necessary safety standards and protocols before allowing further development[10].

This is exactly in accord with the OECD's AI principles[11]: *"AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk".* The onus of proof should be on the developer of novel

---

[2] Foundation models include large language models (such as the GPT series and other generative models, such as Stable Diffusion).
[3] Common Crawl.
[4] GPT-3 (released in 2020) has 175 billion parameters, while the human brain is estimated to have fewer than 100 billion neurons. GTP-4 (released in 2023) has been reported to have one trillion parameters. Note that parameters are not equivalent to neurons, but serve a similar function.
[5] See "2022 Expert Survey on Progress in AI", AI Impacts (6 August 2022) for a recent overview of opinions on the progress and pace of AI development and experts' perception of catastrophic risk.
[6] "The world needs an international agency for artificial intelligence, say two AI experts", Gary Marcus and Anka Reuel, The Economist (18 April 2023).
[7] "Sparks of Artificial General Intelligence: Early experiments with GPT-4", Microsoft Research (22 March 2023): "elucidating the nature and mechanisms of AI systems such as GPT-4 is a formidable challenge that has suddenly become important and urgent" (page 95).
[8] "Emergent Abilities of Large Language Models", Google Research, Stanford University, UNC Chapel Hill, DeepMind (26 October 2022).
[9] Recent projects in this direction include AutoGPT, a tool to make GTP-3 work as an independent agent with only minimal initial direction from the user.
[10] "Pause Giant AI Experiments: An Open Letter", Future of Life Institute (published on 22 March 2023, signed by 27,000+ technologists and other individuals).
[11] "Robustness, security and safety (Principle 1.4)", OECD.AI Policy Observatory, OECD.

systems, not on society. If developers cannot show that their developments are safe, they must not be allowed to proceed.

The “black box” nature of current state-of-the-art AI models, the unpredictable emergence of new capabilities, and the additional “programmatic scaffolding” that makes them behave as independent unmonitored agents, all point to **the need to stop training models larger or more capable than GPT-4 until their safety can be dependably demonstrated**.

A cessation of training any larger foundational models is **easy to implement and enforce**. Currently, only a small number of large corporations train foundational models (primarily Google DeepMind, Meta, Anthropic, Microsoft and OpenAI). Only they have access to the hardware, expertise, and the hundreds of millions of dollars needed to train state-of-the-art models using current methods.

Regulators need to be adequately empowered to detect, investigate and penalise non-compliance. We propose **tough penalties to deter non-compliance** particularly as this prohibition will mostly affect large corporations which have the resources to understand and adjust their business models accordingly. Strong regulation of big tech companies has proven to be a success with the enactment of EU GDPR[12] and the Australian News Media Bargaining Code[13].

One measure that will aid the prohibition is **monitoring the amount of compute used to train foundational models or large language models**. Companies should be subjected to reporting requirements over a certain threshold of computation and AI capability. Furthermore, chip manufacturers (such as Nvidia) need to provide technical means for transparent, public monitoring of industrial-scale compute.

Governments should work to **establish a multilateral process to impose an international moratorium** on large-scale AI capabilities research and development, similar to nuclear non-proliferation treaties.

The Campaign for AI Safety considers the current state of AI capabilities to be at an optimal level where the benefits are maximised and risks are manageable. There are **still great benefits to harness from the current level of capability** for many years to come. But **making state-of-the-art AI smarter is too risky at present** and requires safety research that has not yet taken place.

[12] “Europe's GDPR has accomplished a lot in its infancy”, Katie Collins, CNET (24 May 2019).
[13] “News Media Bargaining Code”, Australian Communications and Media Authority.

2. Impose safety conditions on AI labs and providers

AI labs and providers[14] are capable of making consequential decisions that potentially affect everyone. There are external costs to developing this new technology that AI labs may not take into account. This market failure warrants regulation to address these risks. We propose several conditions on AI labs' and providers' activities to prioritise safety. Such a scheme could be **modelled after licensing requirements in industries such as financial services and healthcare,** which continue to thrive under strong regulation.

Mandate technical AI safety research

We recommend that a licenced AI lab should be under certain conditions:

1. Invest at least 50% of its aggregate research spend on advancement of alignment, reliability, and explainability until regulators can verify there is not a major risk from their activities.
2. At least half of its research staff should be employed in the safety area and not work directly towards advancing capabilities, modality, or model sizes.
3. External safety evaluators should be required to certify safety of new models and incremental advancements.
4. Internal safety committees need to have a legal veto power on deployment of unsafe systems.
5. Members of the safety committee should be public officers and be liable for deployment of unsafe AI systems.

Pre-deployment safety evaluations

Industry is already conducting some safety evaluations of its technology. This should be made a mandatory requirement. The standards of such evaluations need to be adopted at national and international levels.

At first, they can be modelled based on evaluations run by the Alignment Research Center[15] in the US, and they should include a detailed risk analysis that shows that no new risks are introduced from deployment of these systems and that no advancement of AI capability is achieved as a result of their deployment.

Safety committees

Licensees should be required to have safety committees that are similar to internal risk management committees at banks and other financial institutions. Safety committees must be responsible for AI safety of models and their implementations.

---

[14] We define an "AI lab" as an institution that trains foundational models (i.e. large-scale, highly capable models that include Large Language Models and diffusion models). The companies make decisions regarding technical design, training procedures, training datasets, and other aspects of the models. Examples: OpenAI, Anthropic, Meta, Stability AI.

We define an "AI provider" as a company that provides inference services from these models, including via API. The companies make decisions regarding selection, deployment, availability of the models. Examples: OpenAI, Microsoft, Amazon, Hugging Face.

A company can be both a lab and a provider. Examples: OpenAI, Anthropic.

[15] ARC Evals: https://evals.alignment.org/

1. Safety committees are required to certify the safety of new models and need to have legal veto powers over the deployment of unsafe systems.
2. Safety committees are required to certify that no new AI capabilities are added in the state-of-the-art systems.
3. Members of the safety committee are public officers and are liable for deployment of unsafe AI systems.
4. Company directors are not able to overrule directions of safety committees.
5. At least a third of safety committee members are government appointed and remuneration is publicly funded and not tied to the licensee's financial performance.

Periodic safety audits

Periodic safety audits should be mandatory and cover areas of operation of licensees by AI safety certifiers, at least quarterly. Such audits should be modelled after SOC2 Type 2 certification and financial audits. As such, they should audit:

1. Internal controls within the organisation;
2. The process of training and deployment of models;
3. Evaluations of what kinds of outputs models are producing;
4. Internal logic of models (firstly, making sure they are interpretable and explainable, secondly evaluation of safety of the internal logic);
5. Evaluations of potential new capabilities that can be layered on top of these models by third parties;
6. Methods of inference from the model and moderation of outputs.

Importantly, the standards of these audits need to be continuously revised as new threats and capabilities are identified. In order for it to be implemented fast, an ecosystem of safety auditors and standard-settlers should be first mandated by the government and then established by the industry.

Disclosure of training datasets and model characteristics

AI labs and providers should be required to publicly disclose the training datasets, model characteristics, and results of evaluations. This will help build public trust and confidence in the process of development of artificial intelligence and allow the AI ethics and safety community to assess the risks and performance of models with better context.

Capable national regulators

New AI-focussed **government regulators need to be established with strong powers** to:

1. Inspect activities of AI licensees and appoint permanent observers within licensees who will have power to monitor all activities
2. Investigate and prosecute them for breaches of AI laws and regulations
3. Similar to the operation of FDA in the USA, issue consent decrees and suspension orders in relation to unsafe activities in relation to training models, performing inference, API access.
4. Similar to banking prudential regulators, withhold licences and suspend operations altogether until safety can be proven.

3. Redirect existing research funding to AI alignment and safety

Many countries (USA, UK, EU countries in particular) have invested billions of dollars to develop, test and use AI technology. Most (if not all) of the funding has supported the industry. This is in addition to hundreds of billions of dollars of private investment.

Because the AI industry has now reached a critical mass where it is financially self-sustaining, **public funding should now be redirected towards research and development of AI safety protocols and techniques.** This should include the development of AI verification and validation techniques, as well as methods to ensure that AGI systems do not fail catastrophically or cause unintended harm. Governments are in a prime position to support collaborations between companies and researchers and encourage AI labs to embed alignment theories into their models. Specifically, government can redirect existing AI industry support funding towards:

- National standards institutes (such as NIST, ETSI, Standards Australia) to work on means of quantitative assessment of AI capabilities and safety[16],
- Supporting the activities of nonprofits working on AI safety, such as Alignment Research Center or the Center for AI Safety,
- Expanding programmes such as Australian Information Security Evaluation Program to include assessments of catastrophic AI risk from gross misuse.

Universities and departments within them should be given **free choice to use existing funding for AI safety research**, which is currently earmarked for AI capability research, computer science, or fundamental science**.** This way, no new additional outlays will be required from tax-payers, all the while a pipeline of expertise in AI safety could be fostered from new courses and research programmes that focus on AI safety.

[16] "Strengthening U.S. AI Innovation Through an Ambitious Investment in NIST", Antropic (April 2023).