Campaign for AI Safety

Response: AI Accountability Policy Request for Comment April 2023

This submission is made by the Campaign for AI Safety in response to the National Telecommunications and Information Administration's request for comment on AI system accountability measures and policies. We trust that this submission is of assistance to the Department. We are happy to engage further if that would be helpful.

Our main points are:

- *A rules-based, enforceable approach is needed to build public trust and confidence in the development of AI systems and to mitigate the most harmful impacts of AI. In the face of strong competition, companies are not adequately incentivised to develop responsibly or to take voluntary action to reduce the risks of AI systems.*
- *Mandating strong safety conditions and empowering regulators to monitor and enforce are key measures for achieving accountability and thereby help realise the federal objective of advancing trustworthy AI[1].*
- *The onus of proving an AI system's safety should fall on AI labs and providers as they are in a position to prevent the most harms during training and deployment.*

1. Impose safety conditions on AI labs and providers

AI labs and providers[2] 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.

1.1 Mandate technical AI safety research

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

[1] https://crsreports.congress.gov/product/pdf/R/R46795

[2] 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

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.

1.2 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[3] 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.

1.3 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.

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.

1.4 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:

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

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.

1.5 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.

1.6 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.

2. Legislate for joint legal culpability for AI harms

Legislators should pass laws that clarify the **joint legal capability of AI labs, AI providers and parties that employ AI for AI harms**. That means that if the rights of an individual or a customer are infringed, they should be able to sue not just the immediate party that employed the AI system (for example, a small business using AI-powered recruiting software that has a bias against minorities), but also the provider of the AI system (e.g. the recruitment software provider or the API provider such as Amazon or Microsoft) as well as the AI lab that trained and released the AI system (e.g. Anthropic or OpenAI).

Such a joint culpability scheme is needed for three reasons:

1. **Consumers and businesses should have adequate legal recourse** when their rights are violated, including when AI systems are used. The many small businesses that use AI systems to automate their processes will not necessarily have the means to compensate for the damages caused by AI. Including providers and AI labs into the culpability will make sure that damages to individuals and consumers can be compensated by these well-funded entities.
2. **Limiting legal culpability to the businesses that automate their processes with AI would place an inefficient burden on them.** They would either need to themselves invest heavily in safety or seek insurance cover, driving up the cost of doing business.
3. **Law should encourage prevention of harms at the point where it is most addressable.** Prevention of harms is most addressable during training and deployment, therefore it is AI labs and providers who should bear the consequences of misalignment within their systems.

Potential harms to be included in this scheme are:

- **Discrimination and bias.** Example: recruitment software that makes API calls to a large language model provided by a major US technology company includes a bias against people with ethnic surnames and automatically rejects their applications.
- **Copyright violations:** Example: an image generator (similar to Stable Diffusion) has memorised elements of paintings of a certain artist X and adds them freely when a user uses the prompt "in the style of artist X"[4].
- **Privacy law violations:** Example: a large language model was trained on texts scraped from various websites that included inadvertently disclosed private information (e.g. email addresses, physical addresses, dates or birth, etc.). When users prompt it for information, it disclosed the memorised private data.
- **Defamation:** Example: a large language model confabulates stories about individuals, including fake criminal records[5].
- Other areas (such as physical injury) may emerge later as AI technology is implemented more widely.

The provision of joint culpability would not be novel. It would be modelled on existing legal schemes:

- **"Polluter pays"**, which is an internationally accepted principle in environmental law that stipulates that the polluter should pay the damages stemming from the pollution[6]. It is because they are at a point where harms are most easily addressed and prevented.
- **Manufacturer liability for goods with safety defects,** as stipulated in Australian Consumer Law, places the liability provisions for unsafe products on the retailer and the manufacturer, as well as on the company that imported the product from

[4] "Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement", The Verge (7 February 2023)
[5] "Responsibility of OpenAI for defamation by ChatGPT", Michael Douglas, Bennett Litigation and Commercial Law (12 April 2023).
[6] "What is the polluter pays principle?", Grantham Research Institute, London School of Economics (18 July 2022).

overseas, as well as the company that puts itself in the position of of manufacturer (via licensing its brand for the product or promotes itself as the manufacturer[7]).

Both principles are well-established and include common-sense exclusions of liability that ensure that, if placed in an AI context, would not result in a deluge of frivolous lawsuits against AI labs.

About the Campaign for AI Safety

The Campaign for AI Safety is a not-for-profit association with members from around the world. We are an association of people who are concerned about the dangers AI poses to humanity and advocate for a stop on the advancement of AI capabilities and stronger regulation that prioritises safe and responsible AI. We are not affiliated with any political or commercial group. For more information, and to read our policy recommendations, please visit www.campaignforaisafety.org.

[7] "Product liability", Australian Competition and Consumer Commission