Campaign for AI Safety

Consultation: A pro-innovation approach to AI regulation

The Campaign for AI Safety welcomes the opportunity to respond to the Department for Science, Innovation and Technology's white paper, a pro-innovation approach to AI regulation, published in March 2023. We have also participated in the Competition and Markets Authority's AI Foundation Models: Initial review. We trust that this submission is of assistance to the Department.

We have responded in detail to selected questions but our main points are:

- *Stronger, more intelligent AI could act in ways that are catastrophic for humanity[1]. We note that since the white paper's release, the UK Prime Minister, industry and leading AI scientists have expressed concerns about the likelihood of existential risk.*
    - *We urge the Government to prioritise mitigating this risk. The development of more powerful AI should be prohibited and AI labs and providers[2] must be expected to prove their AI is safe and controllable.*
- *Industry supports prescriptive and enforceable regulation[3] (the 'do maximum' option 3 in the white paper's regulation impact statement) and has stated that mitigating the existential risk from AI needs to be prioritised in the same manner as pandemics and nuclear war[4].*
- *We support the white paper's commitment to work towards internationally coordinated action for AI regulation.*
    - *We acknowledge the UK is due to host an AI summit in Autumn this year. We urge the Government to broker a global agreement on AI similar to non-proliferation treaties on nuclear, biological, and chemical weapons.*
- *AI scientists and industry acknowledge the alignment problem is very difficult to solve.*
    - *We recommend redirecting existing industry and research funding into making AI more aligned with the values of UK citizens, specifically the $100 million AI taskforce and the $900 million investment into compute technology in the UK Budget.*

---

[1] AI Risk Center for AI Safety; How Rogue AIs may Arise Yoshua Bengio, 2018 ACM A.M. Turing Award recipient

[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

[3] Sam Altman, CEO of OpenAI, calls for a licensing regulatory framework with mandatory safety standards, pre-deployment safety checks and independent audits: Transcript: Senate Judiciary Subcommittee Hearing on Oversight of AI. Brad Smith, Microsoft President, made similar comments: Microsoft leaps into the AI regulation debate, calling for a new US agency and executive order, CNN.

[4] Statement on AI Risk Center for AI Safety, , which includes signatories from the CEOs of Google DeepMind, Anthropic and OpenAI

CROSS SECTORAL PRINCIPLES

4. How could routes to contestability or redress for AI-related harms be improved, if at all?

New legislation is required to clarify the **joint legal culpability of AI labs, 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). As with products like medicine, UK citizens must have confidence that they are safe. Citizens should not have to fear the AI they interact with is lying, manipulative or harmful.

A joint culpability scheme is needed for three reasons:

- **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.

- **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 (often impractically) in safety or seek insurance cover (expensive due to unknown risk profiles), driving up the cost of doing business.

- **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 certain sub-groups 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”[5].

- **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[6].

---

[5] “Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement”, The Verge (7 February 2023)

[6] Responsibility of OpenAI for defamation by ChatGPT, Michael Douglas, Bennett Litigation and Commercial Law

- Other areas may emerge later as AI is implemented more widely including more powerful and often unpredictable systems. Examples: giving bad advice or faulty code that users don't detect, businesses using ChatGPT API to connect to users' bank accounts, giving rise to risk of automated access and unauthorised withdrawals[7]

The provision of joint culpability could 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[8]. 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 UK 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 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[9]).

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.

5. Do you agree that, when implemented effectively, the revised cross-sectoral principles will cover the risks posed by AI technologies?

The white paper is on the right track in identifying 'safety, security and robustness' as one of the cross-sectoral principles to guide regulation. However, we disagree that this principles-based approach will adequately cover all the risks posed by AI, especially the risk of AI being used or acting in ways that are catastrophic for humanity.

Principles are subject to different interpretations by businesses and regulators and give rise to inconsistent decision-making. We are concerned they are unenforceable, undermine accountability and consumer protections, and ultimately would not lead to a substantial reduction in AI risks and harms.

Both industry and the public expect stronger, more prescriptive regulation. Sam Altman, CEO of Open AI, and Brad Smith, Microsoft President, have publicly called for a dedicated AI regulator, a licensing regulatory framework with mandatory safety standards, pre-deployment safety checks and independent audits. A nationwide survey by the Alan Turing Institute in June 2023 shows 62% of those surveyed want 'laws and regulations that prohibit certain uses of technologies and guide the use of all AI technologies'[10].

In the face of incomplete scientific knowledge and the possibility of serious and irreversible threats of harm, we urge the UK Government to take a preventative approach to directly target the existential risk. While it is true that there currently exists no AI system able to cause an existential catastrophe, the speed of progress towards general capabilities of large language models has taken many by surprise. Industry leaders[11] agree there is a risk to human existence

---

[7] 15 Ways We Are Using ChatGPT in Banking, SouthState Correspondent Division; ChatGPT Use Cases in the Future Banking, Finxtra
[8] What is the polluter pays principle?, Grantham Research Institute, London School of Economics
[9] Product liability" Office for Product Safety and Standards, UK Government
[10] Majority of British public support 'laws and regulations' to guide the use of AI, according to a new nationwide survey, The Alan Turing Institute
[11] Statement on AI Risk, Center for AI Safety, which includes signatories from the CEOs of Google DeepMind, Anthropic and OpenAI

and it should be prioritised and treated on the same scale as the threat of pandemics and nuclear war.

At a minimum, we recommend a regulatory framework that **prohibits further development of more powerful AI systems such as large language models**. Stopping the development of foundation models more powerful than GPT-4 is vital to prevent the development of autonomous agents which behave more intelligently than humans. Ensuring humanity remains the smartest thing on the planet is key to keeping control of our future. In addition:

- Place the burden of proof on the proponent of a proposed activity that there is no harm (or prove safety).
- Make the responsible parties liable for harm (see our response to question 4 on implementation).
- Impose safety conditions on AI suppliers.

Prohibiting further development does not need to be permanent. As with any regulatory approach, policy and legislation can be adapted in response to new information.

Prohibiting AI research in the direction of general capabilities and superintelligence would not prevent innovation of beneficial, narrowly focused AI such as systems that improve medical care or make transport safer (and examples of beneficial AI outlined in the white paper).

We consider 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** and better use of current foundation models for many years to come.

A STATUTORY DUTY TO REGARD

7. Do you agree that introducing a statutory duty on regulators to have due regard to the principles would clarify and strengthen regulators' mandates to implement our principles, while retaining a flexible approach to implementation?

We disagree with introducing a statutory duty on regulators to have due regard to the proposed principles. This approach of **relying on different regulators to use existing laws that are not applicable to AI will not adequately protect UK citizens from the dangers of AI**.The white paper notes some regulators have expressed concerns that they lack the statutory basis to consider the application of the principles.

This approach is limited in protecting UK citizens because, while it compels proactiveness within existing remits, it does not give regulators any new enforcement powers to directly address harms. Instead, the onus is on regulators to "engage with government to explain that their existing legal powers do not allow them to fully enforce compliance"[12].

The regulation impact statement states that compliance with the proposed cross-sectoral principles and accompanying guidance will not be legally mandated. so there is no obligation by businesses to take these into account. We are concerned that unregulated businesses will prioritise 'first mover' advantages with rushed releases at the expense of rigorous testing, resolving safety defects, or validating outcomes for bias.

[12] UK Artificial Intelligence Regulation Impact Assessment page 30

The regulation impact statement states the rationale for preferring to delegate to existing regulators with a duty to regard the proposed principles is to enable a faster launch of the regulatory framework, to allow iterations to keep pace with change, to minimise additional burdens on UK businesses and not impede innovation[13]. We are concerned that **with this 'wait and see' approach, it may be too late to stop societal-scale damage** if leading AI scientists such as Geoffrey Hinton are right about AI posing a risk to humanity's existence[14].

Our recommended regulatory framework (see above in response to question 5) will require new legislation but it will be proactive in mitigating a catastrophe. The prohibition on creating more powerful AI systems effectively targets a limited number of multinational corporations as they have the resources necessary to make training models larger or more capable than GPT-4. These businesses are in a position to absorb compliance costs and ensure responsible development. With this approach, **small businesses and potential new entrants will still have access to current state-of-the-art AI and build products off them. This policy continues to promote innovation that benefits the UK while mitigating severe harms.**

8. Is there an alternative statutory intervention that would be more effective?

Yes, mandatory rules **combined with heavy penalties[15] to deter non-compliance would be more effective**. This will require new legislation.

TOOLS FOR TRUSTWORTHY AI

21. Which non-regulatory tools for trustworthy AI would most help organisations to embed the AI regulation principles into existing business processes?

We agree with using assurance techniques and technical standards to support the UK's regulatory framework. There already exists the UK's AI Standards Hub[16] and the Centre for Data Ethics and Innovation's work on AI assurance techniques for anyone developing AI systems to use[17].

The decisions of AI labs and providers in developing new technology have external costs that potentially affect everyone. As the white paper's regulation impact statement notes[18], these businesses are not incentivised to take these external costs into account.

To address this market failure, we recommend making technical and safety standards mandatory requirements. This is a sure way to address key AI risks and boost public trust to harness the opportunities and benefits that AI technologies present.

Such a scheme could be **modelled after licensing requirements in industries such as financial services and healthcare**, which continue to thrive under strong regulation.

---

[13] UK Artificial Intelligence Regulation Impact Assessment page 33-34

[14] 'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation

[15] We agree with the suggested fines in option 3 of the regulation impact statement ("up to £26.4 million / 6% of global revenue for non-compliance with either unacceptable or high risk requirements, £17.6 million / 4% of global revenue for any other regulatory requirement under the proposal, and £8.8 million / 2% of global revenue for incorrect provision of information relevant to the requirements" page 41).

[16] AI Standards Hub

[17] CDEI portfolio of AI assurance techniques, Centre for Data Ethics and Innovation

[18] UK Artificial Intelligence Regulation Impact Assessment, page 14.

Specifically, we recommend the following:

1. 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[19] 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 dangerous capabilities of AI are achieved as a result of their deployment.

2. 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. We propose the following features:

a) certify the safety of new models, with the legal veto powers over the deployment of unsafe systems;

b) certify that no new potentially dangerous capabilities are added in the state-of-the-art systems;

c) members of the safety committee are public officers and are liable for deployment of unsafe AI systems;

d) company directors are not able to overrule directions of safety committees; and

e) 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.

3. 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:

a) internal controls within the organisation (such as safety training, independence of the safety committees);

b) the process of training and deployment of models;

c) outputs models are producing;

d) internal logic of models (including ensuring that the logic is fully interpretable and explainable);

e) potential new capabilities that can be layered on top of these models by third parties; and

f) methods of inference from the model and moderation of outputs.

---

[19] ARC Evals

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 mandated by the government and established by industry.

4. Disclosure of training datasets and model characteristics

AI labs and providers should be required to publicly disclose the training datasets, model characteristics, and the full results of evaluations (including both positive and negative test results). 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.

FINAL THOUGHTS

22. Do you have any other thoughts on our overall approach? Please include any missed opportunities, flaws, and gaps in our framework.

AI safety research is nascent and **solving the 'alignment problem' is very difficult** (ensuring the 'values' of increasingly powerful AI are aligned with human values), as acknowledged by OpenAI[20] and others[21].

As the UK Prime Minister recently acknowledged[22] far more safety research is needed to solve the alignment and control problems of AI.

We recommend **mandating AI labs to conduct technical AI safety research**:

- Invest at least 50% of its aggregate research spend on advancement of alignment, reliability, and explainability[23] until regulators can verify there is not a major risk from their activities.
- 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.
- External safety evaluators should be required to certify safety of new models and incremental advancements.
- Internal safety committees need to have a legal veto power on deployment of unsafe systems.
- Members of the safety committee should be public officers and be liable for deployment of unsafe AI systems.

Being mostly dominant incumbent businesses with large amounts of resources, we think this is a reasonable cost of doing business. This is also a necessary investment to **build knowledge which will have positive spillover effects on the development of safe AI** (e.g. lower costs for new market entrants).

In addition, we urge the UK Government to **refocus existing industry and research funding into safety and alignment research**, specifically the $100 million AI taskforce and the $900 million

[20] Our approach to alignment research, OpenAI.
[21] What is the AI alignment problem and how can it be solved?, New Scientist.
[22] UK Prime Minister, Rishi Sunak on Investing More in AI Safety Research, 20VC, 14th June 2023.
[23] The other 50% can be spent on optimising foundation models up to GPT-4 capabilities for specific commercial or public applications.

investment into compute technology in the UK Budget. Specifically, we suggest funding towards, government can redirect existing AI industry support funding towards:

- National standards institutes to work on means of quantitative assessment of AI capabilities and safety;[24] and
- Supporting the activities of nonprofits working on AI safety, such as Apollo Research, Alignment Research Center, or the Center for AI Safety.

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. If such discretion is afforded, no new additional outlays will be required from tax-payers, all the while a pipeline of expertise in AI safety could be fostered from adjacent courses and research programmes.

FOUNDATION MODELS AND THE REGULATORY FRAMEWORK

F2. Do you agree that measuring compute provides a potential tool that could be considered as part of the governance of foundation models?

We agree with using compute to govern foundation models. It is excludable (cannot be used by more than 1 player at the same time and can be restricted), quantifiable (therefore detectable), and (currently) required in large amounts for the development of foundation models[25].

Microsoft has suggested a compute-based threshold to monitor powerful new AI models and advanced data centres[26] in the interim until a more complex capability-based threshold solution can be developed. It has also suggested imposing licensing requirements similar to telecommunications and financial services regulation on AI labs (operators of data centres with hardware capable of training large foundation models) to comply with safety redevelopment and deployment requirements.

It will be important to carefully define the compute threshold. As per our policy recommendation to stop the development of ever more powerful AI, the **threshold should not be greater than the compute capacity required to run current foundation models such as GPT-4**. It is easy for businesses to provide this information and for regulators to monitor.

AI SANDBOXES AND TESTBEDS

S1. Which of the sandbox models described in section 3.3.4 would be most likely to support innovation?

We have no comment on the preferred model to support innovation. **Safety must be the number one priority** when designing a regulatory framework. Sandboxes and testbeds may be inadequate given this consideration.

The white paper implies there will be no additional resources available to regulators. We are concerned that, without adequate resourcing, regulators will be exposed to undue influence and reliance on regulatees that have greater technical knowledge.

---

[24] Strengthening U.S. AI Innovation Through an Ambitious Investment in NIST, Anthropic
[25] Presentation on Introduction to Compute Governance, Lennart Heim, Centre for the Governance of AI
[26] Governing AI: A Blueprint for the Future, Microsoft

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**. As the CMA response to this white paper notes, the AI industry is dominated by the most powerful firms that can act as quasi-regulators in the markets they operate[27].

The white paper takes forward Sir Patrick Vallance's recommendation for a multi-regulator AI sandbox to be established. We refer to SIr Patrick Vallance's Regulation for Innovation report which emphasises the **importance of funding and rapid recruitment of specialist talent and skills to establish a high-profile sandbox staffed with AI experts will support the scale-up of businesses and increase investment opportunities**[28].

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.

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

[28] Pro-innovation Regulation of Technologies Review Digital Technologies Report, Sir Patrick Vallance, page 8