Campaign for AI Safety

Request for information: National Priorities for Artificial Intelligence

This submission is made by the Campaign for AI Safety in response to the Office of Science and Technology Policy's request for information to develop a National AI Strategy in the United States. 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 have responded in detail to selected questions but our main points are:

- *A national AI strategy needs regulations that obligate AI labs and providers[1] to develop safe and controllable AI systems. Robust and effective enforcement of these regulations is essential. We propose several conditions on AI labs' and providers' activities to prioritise safety below in this submission.*
- *The development of more powerful and smarter AI needs to be stopped as a matter of priority until it is proven to be safe to continue. Many leading AI researchers believe generative AI can eventually be trained to be far more intelligent than any human and may threaten human values, dignity and existence. Many, including AI labs, warn that the consequences could be as catastrophic as nuclear war or pandemics.*
- *To be dangerous, 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. We make specific recommendations to mitigate misuse of AI.*
- *Our recommendations are directed at the unchecked development of high-risk generative AI (e.g. GPT or Bart). They are not intended to inhibit innovation of beneficial, narrowly focused AI technology.*

[1] 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 (e.g. 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 (e.g. OpenAI, Anthropic).

1. What specific measures – such as standards, regulations, investments, and improved trust and safety practices – are needed to ensure that AI systems are designed, developed, and deployed in a manner that protects people's rights and safety? Which specific entities should develop and implement these measures?

We propose conditions on AI labs' and providers' activities to prioritise safety. These conditions should be legally binding to address key AI risks and boost public trust to harness the opportunities and benefits that AI technologies present.

Specifically, we recommend the following:

Pre-training authorisations

Pre-training model evaluations[2] will require AI companies to state the intended characteristics of AI models, their expected behaviours, and datasets used in training the data. They need to be made publicly available in order to allow the AI safety research community to assess the level of dangers posed by the models. AI safety researchers need to have a mechanism to object unsafe model training and regulators need powers to prohibit creation or fine-tuning of AI systems that are potentially dangerous.

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] and the framework proposed by Google[4], 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.

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;

[2] "The Case for Pre-emptive Authorizations for AI Training", Lennart Heim (10 June 2023).
[3] ARC Evals: https://evals.alignment.org/
[4] "Model evaluation for extreme risks", Toby Shevlane, et al. (24 May 2023).

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.

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.

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.

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.

One way of implementing this is to mandate the provision of model cards[5] to businesses looking to purchase access to AI systems. This is documentation detailing a model's performance characteristics, training data (comprehensive references, not vague descriptions), context in which models are intended to be used, details of the performance evaluation procedures. This will help businesses evaluate the suitability of these systems to their context. This is similar to how the electronic hardware industry provides datasheets with detailed characterisations of components' performances under different test conditions.

[5] "Model Cards for Model Reporting", Margaret Mitchell, et al. (14 Jan 2019).

Mandatory technical AI safety research:

a) Invest at least 50% of its aggregate research spend on advancement of alignment, reliability, and explainability[6] until regulators can verify there is not a major risk from their activities.

b) 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.

c) External safety evaluators should be required to certify safety of new models and incremental advancements.

d) Internal safety committees need to have a legal veto power on deployment of unsafe systems.

e) Members of the safety committee should be public officers and be liable for deployment of unsafe AI systems.

A government body with appropriate enforcement powers and expertise to regulate AI

To implement our recommendations, a scheme could be modelled after licensing requirements in industries such as financial services and healthcare, which continue to thrive and innovate under strong regulation.

This could be a new entity, an extension of an existing department such as the Department of Commerce or an agency such as the National Institute of Standards and Technology. In our view, the speed of implementation should take priority. It may be faster and less costly to make use of an existing Department than to create a new entity. Additional funding would be needed to attract talent and develop monitoring and compliance expertise in AI.

2. How can the principles and practices for identifying and mitigating risks from AI, as outlined in the Blueprint for an AI Bill of Rights and the AI Risk Management Framework, be leveraged most effectively to tackle harms posed by the development and use of specific types of AI systems, such as large language models?

Mandating the principles outlined in the Blueprint would be a more effective way of mitigating the harms posed by AI systems, combined with appropriate enforcement powers and heavy penalties to deter non-compliance (e.g. 6% of global revenue). Legally binding, direct government regulation is needed because there is little incentive for AI labs to prioritise rigorous testing, resolving safety defects, or validating outcomes for bias in the race to achieve ‘first mover’ advantage in the development of more powerful and smarter AI systems.

A voluntary, self-regulation approach would be limited in protecting US citizens from harms because there is no obligation for businesses to take principles and practices such as those outlined in the Blueprint into account.

[6] The other 50% can be spent on optimising foundation models up to GPT-4 capabilities for specific commercial or public applications.

5. How can AI, including large language models, be used to generate and maintain more secure software and hardware, including software code incorporating best practices in design, coding and post deployment vulnerabilities?

Our recommendations in the response to Question 1 will to some extent help ensure the design of more secure and controllable AI pre-deployment.

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

There is very little safety research into AI at present. For example, the alignment problem is at least very challenging to solve and possibly impossible to solve. It is concerning that the current state-of-the-art large language 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.

Until AI labs can demonstrate beyond doubt that their AI systems are safe, we recommend the United States not allow the training of models larger or more capable than GPT-4 (with more compute than 10^23 FLOP[10]). Arresting the development of more powerful and smarter AI is a view shared by leading AI researchers and computer scientists[11].

Stopping further development would only affect a small number of very powerful large corporations

This measure 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.

Prohibiting further development does not need to be permanent. As with any regulatory approach, the government's position 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 beneficial AI that allows us to tackle the world's most important problems.

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.

---

[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] Stop AGI: Proposals: https://www.stop.ai/proposals
[11] Pause Giant AI Experiments: An Open Letter, Future of Life Institute (published on 22 March 2023, signed by 27,000+ technologists and other individuals); "The Godfather of AI' Quits Google and Warns of Danger Ahead", New York Times (1 May 2023).

This recommendation would be even more effective if the United States works with other countries to impose an international moratorium on large-scale AI capabilities research and development (see our response to question 11).

Using compute monitoring for large models

Compute 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[12].

Microsoft has suggested a compute-based threshold to monitor powerful new AI models and advanced data centres[13] 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. 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.

7. What are the national security risks associated with AI? What can be done to mitigate these risks?

We focus our response to this question on the use of AI in managing and operating critical infrastructure (i.e. road and air traffic, utilities, dams, communications and other infrastructure that maintains daily life).

We are most concerned with malfunction in AI or its misuse (e.g. cyberattacks) leading to system-wide failures and ultimately putting the American lives at risk. AI systems in critical infrastructure should be tested for safety and controllability.

11. How can the United States work with international partners, including low- and middle income countries, to ensure that AI advances democratic values and to ensure that potential harms from AI do not disproportionately fall on global populations that have been historically underserved?

An international outlook on regulating AI is paramount because AI technology transcends jurisdictions. Specifically, we urge the United States to work with other countries to develop a legally binding treaty or global agreement to indefinitely stop businesses, governments, and any individual from further development of more powerful and intelligent AI, until at least appropriate international governance and legal structures are in place to ensure the safe and responsible development of AI.

The treaty could be similar to existing non-proliferation treaties on nuclear, biological, and chemical weapons. It could have the following provisions:

[12] "Presentation on Introduction to Compute Governance", Lennart Heim, Centre for the Governance of AI (17 May 2023).
[13] "Governing AI: A Blueprint for the Future", Microsoft (25 May 2023).

- Shutting down large GPU and TPU clusters (the large computer farms where the most powerful AIs are refined).
- Prohibition of training ML models (or combinations of models) greater than 10^23 FLOP in compute (approximately the amount of compute used for the original GPT-3 175B)[14].
- Prohibition of the use of quantum computers in any AI-related activities.
- A general moratorium of large-scale AI capabilities research and development.
- Passing of national laws criminalising the development of any form of Artificial General Intelligence (AGI) or Artificial Superintelligence (ASI).
- Establishment of an international body to oversee the treaty.
- Effective mechanisms for enforcement of the treaty.
- The treaty should not expire until it is universally agreed that it is safe and ethical to resume large-scale AI capabilities research and development.

To aid efforts, the Campaign for AI Safety is running a law student competition to draft a treaty on moratorium of large-scale AI capabilities research and development.

Achieving strong international regulatory cooperation and consistency in AI regulation and knowledge sharing will help protect US consumers and reduce regulatory burden on businesses.

13.How might existing laws and policies be updated to account for inequitable impacts from AI systems?

Generative AI arrived largely unanticipated by the public, creative professionals, publishers and relevant stakeholders. The vast majority of creators never imagined that their works would be used for machine learning. Copyright law never considered this novel use and the ease and speed with which content could be generated with AI. This has created a climate of uncertainty for U.S. creative professionals as well as businesses using AI-generated content.

Therefore, we recommend that laws are clarified and reinforce the copyright regime in relation to training of AI models on copyrighted materials. Specifically:

1. Third parties who wish to use copyrighted materials in training AI models must obtain **specific consent** from the owners of materials.

2. Specific consent implies that third parties **must not be allowed to coerce copyright holders** to provide such consent. For example:

    a. Consent must not be part of terms and conditions of unrelated services (such as video or gaming distribution platforms). This way social media companies will not be able to deny services or features of their platforms to those who do not consent to such use of copyrighted materials.

[14] Stop AGI: Proposals: https://www.stop.ai/proposals

   b. Specificity of consent means that it needs to be a separate agreement, ideally a contract with consideration, which copyright holders have an opportunity to take time to consider, review, and negotiate. It must not be a checkbox underneath a registration form on a website.

3. **Consent must be granular**. For example, copyright owners should be able to specify if they allow third parties to have technical means to generate works "in the style of" the author of the materials.

4. Copyright holders must have effective means of **negotiating and receiving compensation** for the use of their copyrighted materials (such as opt-in collective bargaining mechanisms).

5. **Data used in training AI models must be fully referenced**. References to the works used in training must be made publicly available.

6. Training models using "synthetic data" generated by other models trained on copyrighted materials must be considered equivalent to training on those copyrighted materials.

7. Legal liability in case of infringement of copyright must apply both to parties that train models and parties that use those models to generate content.

8. "Fair use" provisions in law must not apply to AI model training, fine-tuning, or prompting.

9. Materials generated completely or substantially by AI models must not be copyrightable.

These clarifications should be legislated, without waiting for courts to provide them. Appropriate legislation, if promptly enacted, will give certainty in relation to these important questions to the AI industry, the U.S business community, and copyright holders, including artists, writers, publishers, and the wider public.

20. What are potential harms and tradeoffs that might come from leveraging AI across the economy?

We focus on competition and consumer protection harms in response to this question.

The market for advanced AI systems based on large language models is dominated by the Big Tech companies, which are some of the most powerful businesses in the United States and globally. Currently there is strong competition, dynamism and innovation: every few weeks AI labs are releasing new updates or announcing they are building something superior to a competitor's product[15]. We believe the market will evolve to increased concentration of market power and consumer detriment, ultimately harming innovation, productivity, and economic growth. Incumbents may impose higher prices, lower quality. This has been the case for digital platforms which are characterised by vertical integration and durable market power[16].

[15] "Google DeepMind's CEO Says Its Next Algorithm Will Eclipse ChatGPT", Will Knight, Wired (26 June 2023) and "China's Baidu claims its Ernie Bot beats ChatGPT on key tests as A.I. race heats up", Arjun Kharpal, CNBC (27 June 2023).
[16] "Stigler Committee on Digital Platforms: Final Report", the Stigler Center for the Study of the Economy and the State at the University of Chicago Booth (16 September 2019).

Businesses can limit access to large language models to restrict competition in other markets

AI labs currently commercialise large language models by selling access through API but not all are granted access. For example, there is a waitlist for GPT-4 which asks for specific information about intended use. If a player like OpenAI were to retain significant market power, it would be in a position to greatly influence competition in downstream or related markets. AI can exhibit significant economies of scope and be used across multiple markets, this could lead to AI labs acting as de facto conglomerates or quasi-regulators, affecting the ability and terms on which consumers and other businesses can trade. They can make entry by new businesses very difficult, greatly reducing competition in the US. We have seen this played out in digital markets whereby online platform businesses leverage market power into downstream markets using their platform and giving preferential treatment to third party suppliers[17].

Businesses can engage in unfair trading practices such as collusion and personalised social manipulation which reduces competition and diminishes consumer welfare

Algorithms can facilitate tacit collusion and shape consumer behaviour towards outcomes that favour businesses at the expense of competition and consumer welfare[18]. There is a 2015 case involving the pricing of posters in Amazon Marketplace by US and UK sellers promising not to undercut each other[19]. The automated software used in this case seems simple now but it demonstrates the potential for more powerful AI to engage in sophisticated collusion. It may be challenging for regulators to detect and collect evidence to prove AI-generated collusion.

Additionally, there are disturbing examples of digital platforms using algorithms to deceive customers into making choices that benefit the business and engage in discriminatory personalised pricing[20]. Such unfair trading practices erode consumer confidence and trust in the economy and harms innovation and competition.

We believe government intervention is required to stop these businesses from maximising their own private interests at the expense of the broader public interest. They may also operate unfairly in the marketplace, adversely affecting competitors, smaller businesses and the people who work for their competitors. Large language models' generative capabilities can be broadly applied throughout the economy[21]. This could lead to economies of scope in multiple markets which will give rise to competition concerns. There is empirical evidence that AI investment is associated with increased industry concentration, and higher AI adoption and larger gains from AI investments for larger businesses[22]. Competition regulators may need to step in to control pricing (similar to electricity and gas transmission).

---

[17] "Antitrust: Commission opens investigation into possible anti-competitive conduct of Amazon", Press Release, Commissioner Margrethe Vestager, The European Commission (17 July 2019).
[18] "Artificial Intelligence, Algorithmic Pricing and Collusion", Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello (1 April 2019).
[19] United States, 'Plea Agreement', filed in United States of America v David Topkins, CR 15-00201 WHO, 30 April 2015 (Plea Agreement : U.S. v. David Topkins)
[20] "Dark Patterns: Past, Present, and Future: The evolution of tricky user interfaces", Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar (April 2023).
[21] "General-purpose artificial intelligence", Tambiama Madiega, Members' Research Service, European Parliamentary Research Service (EPRS), (March 2023).
[22] "Artificial Intelligence, Firm Growth, and Product Innovation", Tania Babina, Anastassia Fedyk, Alex Xi He, and James Hodson (18 May 2022).

29. Do you have any other comments that you would like to provide to inform the National AI Strategy that are not covered by the questions above?

The National AI Strategy should consider ways to increase safety and alignment research as this is needed to build safer and controllable AI.

Redirect government AI funding to safety and alignment research

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[23] and others.[24]

We urge the United States to refocus existing research funding into safety and alignment research, such as the recently announced $140 million[25] to launch seven new National AI Research Institutes. We also suggest funding in:

- National standards institutes to work on means of quantitative assessment of AI capabilities and safety.[26]
- 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.

Increase remuneration to retain AI knowledge in the public sector and wider society

While scientific knowledge is a public good, the very high wages offered by AI labs to attract talent to privately develop foundation models is depriving the public sector and society of new knowledge.[27] These models are developed without the vital insights being made public, even as patents. We suggest addressing this problem by increasing remuneration to retain talent in the public sector and academia, without having to match the extremely high salaries offered by the big tech businesses. This will allow small players and new entrants access to fundamental knowledge that would otherwise be locked away by the biggest players, a vital precondition for healthy competition. It will also help ease the 'brain drain' from universities and the public sector.

An additional benefit of this investment is that it will build public knowledge which will have positive spillover effects on the development of safe AI (e.g. lower costs for new market entrants).

[23] "Our approach to alignment research", OpenAI (24 August 2022).
[24] "What is the AI alignment problem and how can it be solved?", New Scientist (10 May 2023).
[25] "FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety", The White House (4 May 2023).
[26] "Strengthening U.S. AI Innovation Through an Ambitious Investment in NIST", Anthropic (April 2023).
[27] Samuelson, P.A., 1954. The pure theory of public expenditure. Rev. Econ. Stat. 36 (4), 387–389.