GRADIENT INSTITUTE

Gradient Institute policy recommendations for the AI Safety Summit (1-2 Nov 2023)
Tiberio Caetano and Bill Simpson-Young
27 Oct 2023

AI Safety Summit – Gradient Institute Policy Recommendations

Background on the Summit

What's the aim of the summit? The UK Prime Minister will host the AI Safety Summit 2023 on the 1st and 2nd of November at Bletchley Park, Buckinghamshire, UK. The aim of the summit is to "consider the risks of AI, especially at the frontier of development, and discuss how they can be mitigated through internationally coordinated action." The summit is giving great emphasis to the urgency of the matter: "These risks necessitate an urgent international conversation given the rapid pace at which the technology is developing". The risks in scope (more on them below) are risks to public safety, including risks of major catastrophes.

Can the promised economic benefits of AI be realised without controlling these risks? No. As an analogy, the Chernobyl and Fukushima disasters have devastated the nuclear industry. Moratoriums and phase-outs for nuclear power plants were widespread, and today some nations are grappling with severe consequences. If an AI-triggered catastrophe occurs (such as a cyber-attack bringing down critical infrastructure for an extended time period, or a new and potentially more deadly pandemic), the AI industry will most likely face a crippling backlash and the promised upside would go unrealised.

What types of AI are in scope? The primary focus of the summit is on what's called "frontier AI". These are highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced AI models, and that could possess dangerous capabilities sufficient to pose severe risks to public safety. Frontier models belong to the category of "foundation models", which designate general-purpose AI models (like those driving the ChatGPT web service), as opposed to "narrow AI" models which are built for the purpose of a specific application (such as identifying number-plates in an image of cars or recommending purchases on an e-commerce site).

Does ChatGPT qualify as frontier AI? No. The models driving the ChatGPT web interface (both the free GPT-3.5 model and the paid GPT-4 model) don't have the capabilities of the most advanced AI models in existence, and don't pose the dangers that motivate the summit. These models are foundation models, but are not sufficiently capable or dangerous to qualify as frontier models. The “early” version of GPT-4 however, which is kept in-house at ChatGPT maker OpenAI and is not accessible to users, is a valid example of a frontier AI model given its demonstrated dangerous capabilities disclosed in the official GPT-4 technical report (which includes providing details of how to build a bomb or synthesise dangerous chemicals when asked to do so).

Who has these frontier AI models? Amazon, Anthropic, Alphabet, Inflection AI, Meta, Microsoft, and OpenAI are all known to either possess frontier AI models or to be actively working towards developing them. (These are the same companies that have made a voluntary commitment to the US Administration officials to manage the risks posed by AI.) The open source community has been developing progressively more capable large language models; if that trend continues – and if nothing changes it will – open source frontier models may soon become a reality.

What types of risks are in scope? Only risks that are to a great extent specific to frontier AI models, which happen to be risks that pose a threat to public safety (thus the summit’s title). As stated, these fall within the categories of *misuse* and *loss of control*.

- (Misuse) Example: bad actors such as rogue nations, terrorists, or zealots that use frontier AI to create and proliferate biological or chemical weapons, pandemic-class agents, or cyber-attacks against critical infrastructure
- (Loss of control) Example: advanced AI systems that develop a high degree of autonomy, pursue objectives misaligned with those of humans, and evade our best attempts at controlling and containing their behaviour.

Other AI risks, such as bias, misinformation, discrimination and job losses, are explicitly stated as not being in scope for this summit.

Are these risks speculative? No. The alarm alerting for these risks has been sounded by thousands of AI scientists from industry, government, academia, the nonprofit sector, as well as independent scientists outside the institutional system. These risks stem from the technical reality that dangerous capabilities have evidently surfaced in frontier AI models, and there's no known method to prevent their emergence or completely eliminate them. Certain frontier models have been shown to be capable of facilitating the synthesis of chemical weapons as well as pandemic-class agents. Evidence points to frontier models lowering the barrier for individuals or organisations to conduct cyberattacks, making them more frequent and potentially infrastructure-threatening. There is also a growing body of work indicating that frontier models could evade human control via the emergence of highly advanced, potentially undetectable deception capabilities. Gradient Institute’s submission to the Australian Government on AI regulation provides numerous scientific references substantiating these claims. In conclusion, the concerns that prompted the summit are well-supported by evidence and science.

Why do these risks exist? A key feature of the current AI landscape is that adding more computing power ("compute") to train AI models automatically makes them more capable, whereas safety doesn't automatically follow and requires research and further technology development. In other words, it's easy to make AI smarter (it suffices to use financial capital to buy more high-end AI chips to enable larger model training runs), but hard to make it safe (it requires human capital, research and innovation). Since the demand for intelligence is high, and the supply of AI chips is unregulated, this tends to create a growing gap between AI capability and AI safety.

Why is there urgency in controlling these risks? Existing frontier AI models are proven to be capable of helping non-experts cause a pandemic or create bio and chemical weapons. The capabilities of frontier AI models increase at a rate that safety guarantees do not, and there is great demand and supply for capabilities. The data centres of frontier AI labs, which host these models, do not possess military-grade security. The factors of production required to build frontier AI models are all either publicly available (published algorithms and freely accessible data on the internet) or legally available for purchase on the open market (high-end AI chips). Open-source foundation models are proliferating and their capability levels are increasing and approaching the threshold beyond which they would qualify as frontier AI models. One frontier AI lab CEO stated that models trained with 10x more compute than today's most powerful models will appear in 2024, and 100x in 2025-26. When all this is put together the urgency of the matter becomes clear.

Will current AI standards efforts address the risks? No. They are not focussed on the risks of frontier AI models and the timelines for typical international standards development through existing international standards bodies does not operate at the time scales necessary.

Some frontier AI labs have influenced this summit. Is there a risk of regulatory capture? Yes. Some frontier AI labs are known to have influenced the summit (e.g. OpenAI, Google Deepmind, Anthropic). This is both understandable and desirable, since they have direct insight into the most powerful frontier models, which they have themselves built. However, this poses a risk that they might act in concert to steer the summit towards a path that both entrenches their competitive advantage and sidesteps accountability for harms that may originate from their own frontier models. The summit is conducting a broader consultation with the scientific community, civil society representatives and businesses, which may help control this risk. However, the risk persists. It is paramount that the word "liability" features prominently in the summit discussions, not only in the context of applying frontier AI models but, crucially, in the context of their *development, deployment* and *release* (including open-source release).

Recommendations for all summit participants

In Gradient Institute's view, the summit should aim to achieve the following objectives:

- **Shared sense of urgency for an international agreement**. A shared understanding of the urgency of achieving an international agreement between major nations, crucially including both China and the US, on a minimum viable proposal for effectively managing the risks of misuse and loss of control of frontier AI models.
- **Post-summit working group**. Establishment of a post-summit working group with a mandate to respond to that sense of urgency, aiming at rapidly developing a blueprint for a lean international governance body for frontier AI. The working group should be appropriately resourced with technical AI experts who specialise in frontier AI models. Among these experts should be representatives from major frontier AI labs, but crucially also independent experts as well as experts who are advocates for open-source AI models, so as to mitigate the risk of regulatory capture. The international governance body should be charged with
  - developing novel and minimum international standards specifically for the safety of frontier AI that, if complied with by each separate jurisdiction, will effectively control the risks of frontier AI
  - certifying that the regulatory regimes adopted within individual jurisdictions meet the international standards

Recommendations for Australia

It is our view that Australia should take a leadership role in the summit:

- **Australian leadership during and after the summit.** Australia has significant scientific expertise in AI, but we are not considered to have 'skin in the game' commercially in the same way as the US, China, the EU or the UK. This reinforces our already well established reputation as an honest broker in this field internationally (including by the US and China); it was Australian diplomats that brokered key agreements in 2013 and 2021 on the application on International Law and International Humanitarian Law in cyberspace. We have politicians that understand the issues and are leading the world in considered domestic regulation. We have experts (technical and diplomatic) with established relationships and the ear of great powers. Few countries have this combination; it behoves us not to squander it.