GRADIENT INSTITUTE

Safe and Responsible AI in Australia

Gradient Institute submission in response to the Department of Industry, Science and Resources discussion paper

26 July 2023

Lead Authors:
Lachlan McCalman, Tiberio Caetano

Contributing Authors:
Kimberlee Weatherall, Chris Dolman, José-Miguel Bello y Villarino, Bill Simpson-Young

About Gradient Institute

We are an independent, nonprofit research institute that works to build safety, ethics, accountability and transparency into AI systems. We research new algorithms, provide training and auditing services for organisations operating AI systems, and provide technical guidance for AI policy development. Contact us at info@gradientinstitute.org.

Submission summary

Gradient Institute is grateful for the opportunity to respond to the Department of Industry, Science and Resource's consultation on Safe and Responsible AI in Australia.

This response highlights two key issues in the discussion paper:

1. The 'risk-based' approach to AI regulation outlined fails to properly target controls towards the context-specific risks posed by an application and would lead to ineffective management of those risks.
2. The new risks to public safety created by certain highly capable foundation models ('frontier' models) are not acknowledged or addressed in the proposed approach.

To address these issues, we recommend the government take the following steps:

1. Treat application-specific risks through existing sector-specific and general regulation by ensuring that existing regulation is being applied, providing guidance on how to apply it and updating or clarifying it to treat new AI-associated risks as needed.
2. Lead a global regulatory response to the safety risks of frontier models by investing in the development of standards and compliance mechanisms, championing international agreements to apply them, and implementing them in Australia.
3. Create a government body with access to technical expertise that can assist existing regulators with their AI response, create and potentially enforce regulation for frontier model development, and advise the government on the rapidly changing AI technology landscape.

Introduction

Gradient Institute believes that design of AI regulation should be motivated by careful examination of the risks AI creates. However, for such an approach to be effective, it must

1. identify the key risks arising from use and development of AI
2. determine the mechanisms causing those risks to arise
3. provide effective risk controls targeting those mechanisms.

In this response, we define two distinct categories of AI risk caused by different mechanisms that are relevant to the discussion paper:

- **AI application risk:** risk arising from the use of AI in a particular **application,** present because of an interaction between properties of the application and properties of the AI system
- **AI development risk**: risk arising from **AI development,** present before the AI system is applied to a particular domain.

We argue that targeting these different mechanisms requires three classes of regulation:

- **sector-specific regulation** such as professional standards for medical practitioners or requirements for vehicle safety, which helps control AI application risk
- **general regulation,** such as privacy or anti-discrimination law, which also helps control AI application risk (as well as other non-technology-specific risks such as copyright violation), but applies similar interventions across application boundaries
- **technology-specific regulation**, analogous to that which governs the creation of steam boilers or recombinant DNA products, which we claim will be required to help control AI development risk.

Based on these observations, we contend:

1. The 'risk-based approach' of preset risk-tiering outlined in the discussion paper (especially for instance Box 4) wrongly tries to control application risks across a wide range of contexts with technology-specific regulation. This approach fails to properly target controls towards the context-specific risks posed by an application and will lead to ineffective management of those risks. What is needed instead is support for

existing sector-specific and general regulations to be clarified, modified or enhanced as needed to respond to the use of AI in their area of concern.

2. The new safety risks created by certain highly capable foundation models ('frontier' models) are not acknowledged. These risks arise from AI development and are not specific to an application sector. New technology-specific regulation of the development of frontier models is required in order to prevent these risks from arising.

The following two sections expand on these points. We believe regulation of AI application risk is well addressed by other submissions and refer the reader to those from the Australian Institute of Actuaries[1] and the ARC Centre of Excellence on Automated Decision-Making Society[2]. Section 1 therefore provides a brief overview of our concerns with the discussion paper's 'risk-based' approach and our proposed alternative, asking the reader to refer to these submissions for more detailed discussion. Section 2 is the main focus of our submission, as we believe the safety risks arising from the development of frontier AI models are not yet widely appreciated in Australia.

1. Controlling AI application risk requires sector-specific regulation

This section is relevant to questions 2, 4, 8, 14 and 15 in the discussion paper.

Preset risk-tiering outlined in the paper (especially for instance Box 4) wrongly tries to control application risks across a wide range of contexts with technology-specific regulation. The approach relies on two key steps:

1. assessing the risks of an AI system by using broad descriptions of 'impact' that map into risk tiers

2. providing a preset list of general risk controls that should be applied at each tier.

These steps are defined at a general level across the wide gamut of applications possible for AI. The proposed approach will be ineffective for two reasons:

[1] https://www.actuaries.asn.au/Library/Submissions/2023/2320725SubDoISRAI.pdf
[2] The ADM+S submission was provided to us in draft form.

- **Risk is not one-dimensional.** For example, a system may have a high risk of causing harm by unfairly discriminating, but a low risk of causing harm by spreading misinformation.
- **Controls are not universally effective across different risks.** Whether a particular intervention is effective at controlling a risk is determined by the details of the application and the specific risk it targets, not a system's overall 'risk level'.

The result will therefore be:

- **ineffective control of some risks:** some systems will have risks not effectively controlled by the general interventions imposed due to their 'risk level'
- **needlessly onerous controls:** some interventions imposed due to the system's assigned risk level may be costly to implement but not effectively control the risks associated with that system.

The approach suggested in the discussion paper also breaks down when many of the harms it aims to control are not AI-specific, but rather could equally be caused by other technological mechanisms. This leads to:

- **inconsistent and duplicated regulation:** systems using AI must potentially comply with both sector-specific regulation and AI regulation, creating problems if those two regulations overlap or conflict
- **loopholes:** developers may avoid complying with AI regulation by modifying their technology to fall outside the provided definition of 'AI' (a label that resists neat categorisations, with most experts not agreeing on a precise definition).

Consistent with the submissions by the Australian Actuaries Institute and the ARC Centre of Excellence in Automated Decision-Making and Society, we believe that controlling AI *application* risks is best achieved through outcome-focused regulation. Given that the outcomes of concern are likely to be addressed by existing regulation of that application, this suggests a focus on applying, and if necessary, updating sector-specific and general purpose regulation.

Recommendation 1: Within the scope of each sector-specific and general regulation, we recommend the Government:

1. ensures that existing regulation is being applied to AI systems
2. assesses whether regulators or practitioners require guidance or clarification for how existing regulation should be applied to AI systems
3. determines whether the use of AI creates new risks, identifies those risks specifically, and updates the regulation with appropriate controls as required.

Recommendation 2: We recommend the government forms a body with access to appropriate technical and non-technical expertise in AI and automated decision-making to support sector and general regulators in taking these actions.

Such a body would alleviate the need for existing regulators from each having to acquire that expertise themselves, and could also serve as a general advisory body for government on this fast-moving and consequential technology.

Priority areas for update of general purpose regulation include privacy/data protection law, anti-discrimination law, and consumer protection law. In particular, there is a need to ensure that software-driven and data-driven (including AI-driven) goods and services meet basic quality standards. Consumer protection regulation could potentially impose basic quality standards similar to those in the draft EU-legislation (for example, relating to data and model quality, transparency and reliability), and addressing these important questions via general purpose regulation would ensure they cover other software and automation beyond AI.

This is not to say there is no role for technology-specific regulation of AI. Technology-specific regulation is suitable for controlling risks arising from the process of AI development itself, and for which the details of the application are not relevant to the design of effective controls. The safety danger from the development of frontier models is the most important risk of this type, and must be addressed urgently with technology-specific regulation.

2. The development of frontier AI models needs regulation

This section is relevant to questions 2, 4, 8, 10, 14 and 19 in the discussion paper.

‘The extent of potential risks from advances in AI, such as generative AI, remains uncertain. However, with the rapid acceleration of the development of AI applications, such as ChatGPT, and indications of increased capability, it is time for Australia to consider whether further action is required to manage potential risks while continuing to foster uptake.’

– Safe and Responsible AI in Australia, Discussion paper (page 26)

2.1 Risks from narrow AI

Regulating AI with a combination of sector-specific and general regulation is effective at addressing risks arising from *narrow AI* models. Narrow AI models perform a single specific task such as identifying number-plates in an image of cars or recommending purchases on an e-commerce site, and cannot be applied beyond that application. Risks arising from the use of narrow AI, therefore, exist within the scope of the application for which they are designed.

The risks of a narrow face recognition system, for instance, are related to concerns such as the legitimacy of the reasons for deploying it and the performance of the system across a number of metrics relevant for the application (e.g. overall accuracy, accuracy for different demographic groups, etc.). The risks of a credit card fraud detection system are associated with outcomes like excessive false positives leading to poor customer experience, or excessive false negatives leading to financial loss for the bank. However, a face recognition system can’t suggest wrong explanations for a blood test result and a credit card fraud detection system can’t recommend an inappropriate diet plan.

As narrow AI systems are limited to creating or amplifying application risks, the approach to regulation proposed by the previous section is likely sufficient to control those risks. This involves applying (and as necessary clarifying and modifying) existing sector-specific regulation and applicable general regulations (e.g. consumer protection, privacy, etc.)

2.2 Frontier models

In light of recent AI breakthroughs it has become increasingly clear that an application-focused approach may not be sufficient to satisfactorily address risks posed by *frontier AI models*.

The term 'Frontier AI models' designates 'highly capable foundation models that *could* have dangerous capabilities sufficient to pose severe risks to public safety and global security'.[3] [4] Not all foundation models will fall into this category; only sufficiently capable ones. For instance, GPT-3 almost certainly falls outside the category of a frontier model, whereas the 'early' (not publicly released) version of OpenAI's GPT-4 almost certainly falls within it.[5]

2.3 Risks to public safety from frontier models

The word *capabilities,* highlighted in the quote opening this section, is important in the context of foundation models. Capabilities are the functions that the language or code produced by the model are capable of expressing or carrying out. For instance, out of OpenAI's publicly available models, GPT-4 is more capable than GPT-3.5 because it can express and carry out a broader set of functions, and do so more competently.[6]

In the current state-of-the-art paradigm for developing foundation models, capabilities are not explicitly designed into the model but rather **emerge suddenly and unpredictably** as a byproduct of increasing the computational power (compute) used to train the model.[7] For instance, by increasing training compute by a few orders of magnitude, the resulting model may acquire the capability of unscrambling words, communicating in Farsi or performing modular arithmetic.[8]

This technical fact alone is at the core of the generative AI revolution: increasing the model's performance is not nearly as much about strategic investment or innovative research than it is about resource scaling. Simply procuring more specialised AI chips for training is sufficient to keep enhancing model capability.[9] [10] [11]

---

[3] https://arxiv.org/abs/2307.03718
[4] Emphasis added by the authors of this response. A frontier model won't *necessarily* possess dangerous capabilities.
[5] https://arxiv.org/abs/2303.08774
[6] https://openai.com/blog/chatgpt
[7] https://arxiv.org/abs/2206.07682
[8] https://arxiv.org/abs/2304.00612
[9] https://arxiv.org/abs/2001.08361
[10] Rich Sutton. The Bitter Lesson. Mar. 13, 2019. URL: https://perma.cc/N9TY-DH22
[11] https://arxiv.org/abs/2202.05924

Not all capabilities however are as innocuous as word unscrambling. **Dangerous emergent capabilities are starting to be detected by researchers**.[12] For instance, LLMs are already capable of facilitating the synthesis of chemical weapons[13] as well as pandemic-class agents.[14] There is research indicating that as LLMs become more capable they could potentially develop capabilities for dramatically advanced forms of persuasion,[15] [16] effectively enabling widespread manipulation of individuals for commercial, political, fraudulent or criminal purposes. Evidence is increasing that LLMs could also lower the barrier for individuals or organisations to conduct cyberattacks, making them more frequent and potentially infrastructure-threatening.[17] [18] There is also growing theoretical evidence that frontier models could evade human control via the emergence of highly advanced, potentially undetectable deception capabilities.[19] [20] [21] [22]

Crucially, the fact that a dangerous capability isn't detected in a model doesn't imply it isn't present in the model. The set of capabilities developed by a model isn't overtly legible during or even after the model's development. There is ample evidence of dangerous capabilities being discovered after deployment, which includes the examples just mentioned. Moreover, the impact of dangerous capabilities can be greatly amplified post-deployment, for instance through LLM-integrated applications performing indirect prompt injection attacks.[23]

More generally, there is currently no guarantee that a sufficiently capable frontier model won't possess a highly destructive capability that, even despite best efforts, remains undetected prior to deployment. Because it's impossible to foresee how the model will respond to an entirely new prompt sequence,[24] as long as a dangerous capability *exists* within the model, then *there is a risk* it will eventually be manifested when the model is used.[25]

---

[12] By dangerous we mean sufficient to cause death or significant and potentially irreparable harm to people's wellbeing, dignity or autonomy.
[13] https://arxiv.org/abs/2304.05332
[14] https://arxiv.org/abs/2306.03809
[15] https://arxiv.org/abs/2303.08721
[16] They can already help create disinformation campaigns: https://arxiv.org/abs/2301.04246
[17] https://arxiv.org/abs/2305.06972
[18] https://arxiv.org/abs/2306.12001
[19] https://arxiv.org/abs/2209.00626
[20] https://onlinelibrary.wiley.com/doi/10.1002/aaai.12064
[21] https://arxiv.org/abs/2206.05862
[22] https://arxiv.org/abs/2206.13353
[23] https://arxiv.org/abs/2302.12173
[24] If that was known, there would be no need to use the model.
[25] Following the previous footnote, this risk could in theory be eliminated at the use rather than development stage by constraining use to only include prompt sequences already tested and verified as not triggering dangerous capabilities. However this is technically infeasible for all but sufficiently short prompt sequences composed of sufficiently small prompts – and in this case the model would effectively degenerate into behaving as a lookup table, which would defeat the very purpose of building the model in the first place.

In summary, with the current state-of-the-art approach to develop frontier models:

- we know that as models scale, they develop potentially dangerous capabilities
- we can't foresee what these capabilities will be
- we can’t ensure we’ll identify them prior to deployment.

This implies there's a risk of these dangerous capabilities eventually making their way to real people (including malicious actors) and causing widespread harm.

2.4 Source of frontier model risk

The research reviewed in the previous section shows that frontier AI models differ from narrow AI models in a key aspect that directly relates to risk. For frontier AI models, there is a new risk that is *inherent to model development*, which is separate from any risks that may be associated with the context in which the model is expected to be used. Prompting a model in a certain way may simply ‘awaken’ a dormant dangerous capability already built into the model.

If we want to allow foundation models to be widely deployed, democratised, and used for the prosperity of humanity, we must be sure that such widespread usage won’t eventually lead to the injection of a sequence of prompts that unlocks a dangerous, destructive capability built in during the development stage. The only way to guarantee this is to ensure that no dangerous capabilities exist in the first place, and this can only be done during *development* of the model.

2.5 Regulating frontier model development

Risks are most effectively treated at their source. This helps explain why the development of potentially dangerous technologies — and not only their use when they touch real people — is in general regulated. In Australia, this is true for many potentially dangerous technologies,

including pharmaceuticals,[26] nuclear technology,[27] biotechnology,[28] food safety,[29] and electricity.[30]

Similarly, given the risks intrinsic to the development of frontier models, we believe Australia should take steps to put appropriate guardrails on their development so as to ensure they are safe for deployment. The view that frontier models must be regulated as they are developed, not only as and when they are used, is shared by many AI technology experts.[31]

This does not mean that the development of all frontier models should be banned. Frontier models *could* possess dangerous capabilities, but don't necessarily. Different capabilities will have different degrees of danger. Instead, we believe the onus should be on frontier model developers to demonstrate that their models pose no danger to public safety if released.

This is entirely consistent with the approach already taken with other potentially dangerous technologies: drugs, foods, genetically modified organisms, etc. must by law be shown to be safe before they are released: frontier AI models should be added to the list. For that to happen, a clear line must be drawn between acceptable and unacceptable frontier AI models, and mechanisms to verify and enforce compliance must be created.

Crucially, since the harms potentially arising from frontier AI model development need not respect jurisdictional boundaries (e.g. devastating cyberwarfare or pandemics), an exclusively national approach isn't sufficient to appropriately manage the risk. Countries will need to coordinate and work towards a common baseline for an international approach, possibly involving binding agreements.

Binding agreements addressing life-threatening technologies which nonetheless have a practical use are not unprecedented. The Montreal protocol, which regulates production and

[26] The TGA agency mitigates risk at its origin by enforcing rigorous preclinical and clinical trials. These ensure a potential drug's safety and effectiveness before it reaches the public.
[27] The ARPANSA agency manages risk at the source by enforcing safety standards for handling radioactive material and construction and operation of nuclear facilities, thereby preventing accidents.
[28] The OGTR agency controls risk at its origin by setting lab safety standards and requiring risk management plans before work on new GMOs begins, ensuring that potential risks are identified and managed early
[29] The FSANZ agency manages risk at the source by setting food safety standards. This includes regulations for handling, preparation, storage, and transportation of food products. They ensure the food people consume is safe, correctly labelled and free from contaminants.
[30] The AER agency controls risk at the source by enforcing safety and reliability standards for the generation, distribution, and retail of energy. This includes ensuring power plants meet certain operational and safety standards, electricity networks maintain their infrastructure properly, and electricity retailers comply with market rules. This helps ensure a reliable and safe supply of electricity to consumers and reduces risks such as power outages, fires, and other safety hazards.

[31] https://arxiv.org/abs/2307.03718

consumption of ozone depleting substances, is an example of a highly successful treaty.[32] The treaty on non-proliferation of nuclear weapons permits civilian nuclear technology but limits the development of nuclear weapon technologies by non-nuclear-weapon states.[33]

We believe the following actions, undertaken by the global community, would form the basis of an effective global regulatory approach for controlling the public safety risk posed by the development of frontier AI models:

- **Establish global safety standards** for the development of frontier models.[34] The standards should establish clear safety criteria that the development of a frontier AI model should meet, effectively drawing a line between acceptable and unacceptable models. The standards could then specify different provisions and controls for models that fall within the acceptable risk threshold but otherwise differ in their assessed risk levels. The standards should also be adaptable and capable of swift modifications in response to new evidence.[35]
- Establish mechanisms to **verify compliance** with these safety standards, including transparency requirements such as registration and disclosure, as well as technical mechanisms.[36]
- Establish mechanisms to help **enforce compliance** with these safety standards. This could involve nation-level licensing and liability regimes.
- Ensure that the standards developed and regulatory measures taken **do not stifle innovation and competition** which would make it easier for frontier AI lab incumbents to further cement their advantage.[37]

2.6 Recommendations for Australia's response

Australia has an opportunity to lead international efforts in regulating the development of frontier models.

[32] https://www.unep.org/ozonaction/who-we-are/about-montreal-protocol
[33] https://www.iaea.org/topics/non-proliferation-treaty
[34] Foundation models are well-defined, but determining whether a foundation model qualifies as a frontier model is more ambiguous and requires effort. This suggests the process of defining a frontier model should be part of the standards, and therefore that at least some elements of the standards should be broadly applicable to foundation models in general.
[35] This is in line with recent proposals, e.g. https://arxiv.org/abs/2307.03718
[36] Researchers are starting to develop technical mechanisms to verify compliance with rules for training large language models, see e.g. https://arxiv.org/abs/2303.11341
[37] See https://www.fast.ai/posts/2023-11-07-dislightenment.html for concerns about the risk of concentrating power.

Recommendation 3: We recommend the Government take the following steps to address the risks arising from frontier model development:

1. Invest in the development of safety standards and compliance verification mechanisms for frontier models and make them available to be used as global tools.
2. Lead an international coordination effort on standards and compliance for frontier model development[38], for example by sponsoring and promoting a UN General Assembly resolution.[39]
3. Implement national-level standards, verification mechanisms and enforcement mechanisms compatible with a globally coordinated response.
4. Convene an appropriately resourced body of experts to support the Government in these efforts, charged with staying informed on the latest technical advancements of AI and their implications for public safety, and advising the Government accordingly. This body requires expertise that overlaps with the body proposed in Recommendation 2 for supporting existing regulators and could be combined with it.

[38] Verification of compliance with international agreements in the case of AI may be easier than in the case of nuclear technology: https://arxiv.org/abs/2304.04123
[39] https://en.wikipedia.org/wiki/United_Nations_General_Assembly_resolution