Safe and Responsible AI in Australia Response

Definitions

1. Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?

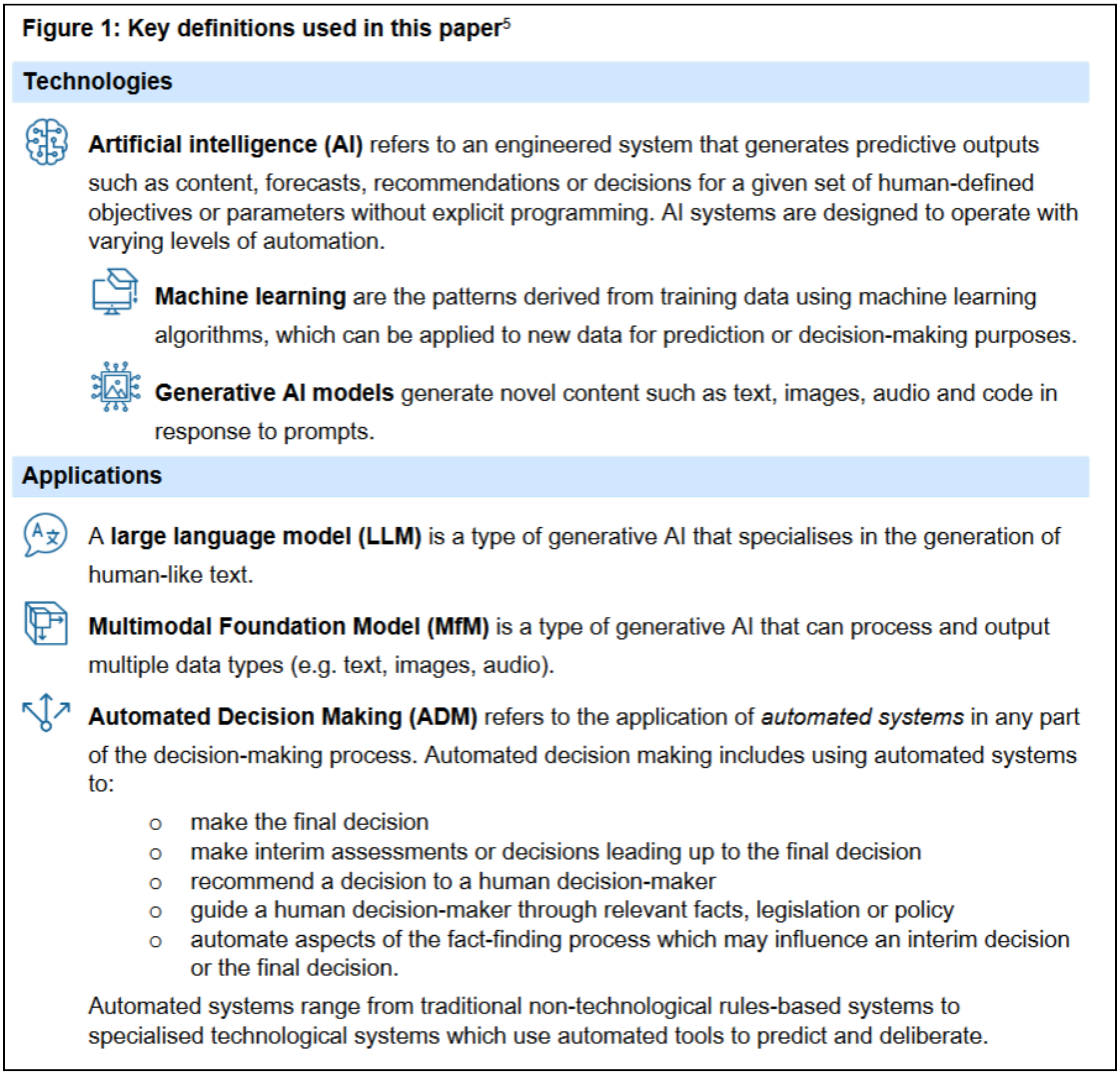

Figure 1: Key definitions used in this paper[5]

Technologies

Artificial intelligence (AI) refers to an engineered system that generates predictive outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives or parameters without explicit programming. AI systems are designed to operate with varying levels of automation.

Machine learning are the patterns derived from training data using machine learning algorithms, which can be applied to new data for prediction or decision-making purposes.

Generative AI models generate novel content such as text, images, audio and code in response to prompts.

Applications

A **large language model (LLM)** is a type of generative AI that specialises in the generation of human-like text.

Multimodal Foundation Model (MfM) is a type of generative AI that can process and output multiple data types (e.g. text, images, audio).

Automated Decision Making (ADM) refers to the application of *automated systems* in any part of the decision-making process. Automated decision making includes using automated systems to:

- make the final decision
- make interim assessments or decisions leading up to the final decision
- recommend a decision to a human decision-maker
- guide a human decision-maker through relevant facts, legislation or policy
- automate aspects of the fact-finding process which may influence an interim decision or the final decision.

Automated systems range from traditional non-technological rules-based systems to specialised technological systems which use automated tools to predict and deliberate.

Feedback on listed terms:

Some AIs take actions, rather than just making predictions, eg RL agents (though there is no sharp line between predicting and acting). This could be delineated by an “agency” axis measuring the capability for independent goal–directed actions.

The “Machine learning” definition seems unnecessarily muddled (unimportant)

Other important terms:

Size:

How many parameters in the AI system.

Compute:

How many floating-point-operations (FLOP) were used in training the AI system.

Multimodality:

The types of data an AI system can handle as inputs.

Narrowness vs Generality:

How specialised the capabilities of the AI system are to the domain in which it was trained, vs how much its capabilities generalise to new domains (even unexpected ones).

Capabilities robustness:

How capably the AI system acts when generalising outside of its training distribution. A general AI system has high capabilities robustness, whereas a narrow AI system has low capabilities robustness.

Value robustness:

How well the AI system adheres to intended values and goals when generalising outside of its training distribution.

Agency:

More ``agentic’’ AIs behave less like tools and more like independent goal-pursuers. There will likely be economic pressure to make systems more agentic, but this may become extremely dangerous due to goal (reward-function or loss-function) misspecification or misgeneralization.

Outer misalignment:

The goal-function (reward-function or loss-function) is misspecified during training, and does not capture what the designers intended.

Inner misalignment:

The goal-function (reward-function or loss-function) is correct during training, but has multiple compatible generalisations out of the training distribution, so the AI system capably pursues unintended goals out of the training distribution. (Capabilities robustness holds, but value robustness does not)

Mechanistic interpretability:

Explaining and predicting the behaviour of AI systems by reverse-engineering the computations they are performing into understandable code.

Scalable oversight:

Designing ways of supervising, updating, and modifying AI systems which will continue to work as these AI systems become capable enough to give false impressions of correct behaviour.

Dangerous capabilities evaluations:

Testing whether AI systems have capabilities which would make them dangerous if they were misaligned or misused, eg. the ability to fabricate convincing falsehoods, the ability to develop potentially harmful new technologies, the ability to formulate or execute long-term plans, the ability to exfiltrate their weights, self-replicate, or self-modify. https://arxiv.org/abs/2305.15324

Alignment evaluations:

Testing whether AI systems have internalised the correct goals and will have value-robustness stronger than their capabilities-robustness out of distribution. https://arxiv.org/abs/2305.15324

Potential gaps in approaches

2. What potential risks from AI are not covered by Australia's existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?

Existential risks:

(extremely high impact risks, on a level such as those posed by nuclear war)

- The statement *"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."* has recently been signed by the heads of major generative AI labs (OpenAI, Google DeepMind, Anthropic, Stability AI) as well as many key academics pivotal in developing this technology (Geoffrey Hinton, Yoshua Bengio, Stuart Russell, etc.) and other notable figures such as Bill Gates. https://www.safe.ai/statement-on-ai-risk#open-letter
- I believe these risks come not primarily from misuse, but from the (potentially inadvertent) creation of highly capable goal-directed AI systems with unintended goals.

- I believe the primary danger lies in the training of extremely large-scale AI systems. Though this mainly happens outside of Australia, Australia's policies on this topic can still have a global impact.

Basic explanation for why I worry about these risks:
AIs are created by repeatedly shifting billions of parameters until they produce good behaviour on their training data. This process gives virtually no understanding of the internal mechanisms they implement - when we look inside these AIs, we cannot usually understand how or why they show a particular behaviour. This may be very problematic as more independently capable "goal-directed" AI systems are created, as seems likely due to market pressures. There are many examples of current AI systems seeming to internalise the wrong goals due to correlated proxies in the training distribution (https://www.deepmind.com/blog/how-undesired-goals-can-arise-with-correct-rewards), but when it occurs in more generally capable AI systems it could be catastrophic. We do not currently have the tools or understanding to look inside these systems and see what "goals" they have actually internalised. So if we can only infer an AI's goals from its behaviour, then we may not be able to act until it is too late. If an extremely capable system internalised unintended goals, it would be instrumentally incentivised not to allow us to change it or discover its misalignment, because it would then not achieve these goals.

In jargon: I'm worried about inner misalignment caused by goal-misgeneralization in general agentic systems, leading to capabilities robustness which far outstrips value robustness, causing catastrophic outcomes due to instrumental convergence.

Possible regulatory action:

- Much more funding and support for technical AI safety research in Australia. Technical solutions to these problems (eg. interpretability, scalable oversight) found in Australia can have a huge impact on the safety of these systems globally. These technical problems must be solved before the technical problems involved in training extremely capable AI systems.
- International cooperation on a system of compute governance or licencing affecting only the largest and most dangerous AI training runs
- The suggestions laid out in *"Frontier AI Regulation: Managing Emerging Risks to Public Safety"* https://arxiv.org/abs/2307.03718 and *"Towards best practices in AGI safety and governance: A survey of expert opinion"* https://arxiv.org/abs/2305.07153 provide excellent templates from which to approach this problem.

3. Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or Impacts.

Much more funding and support should be provided for technical AI safety research in Australia. Technical solutions to these problems (eg. interpretability, scalable oversight, dangerous capabilities detection, goal-directedness detection) found in Australia can have a huge impact on the safety of these systems globally. These technical problems must be solved before the technical problems involved in training extremely capable AI systems.
Funding should be available not only for established academics, but for new researchers entering this new research field, and for new AI safety research labs.

4. Do you have suggestions on coordination of AI governance across government? Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia.

The arguments for existential risks from AI are many, and many of them can be easily misunderstood. It is important that people in government become literate on these topics in order for communication and policymaking to be effective. Some important to understand terms are

- Narrowness vs Generality
- Capabilities robustness vs value robustness
- Goal misgeneralisation
- Inner misalignment vs outer misalignment
- Instrumental convergence

Responses suitable for Australia

5. Are there any governance measures being taken or considered by other countries (including any not discussed in this paper) that are relevant, adaptable and desirable for Australia?

Australia must avoid the mistakes of previous AI regulation attempts, such as assuming that the biggest risks come only when applying AI to specific high-risk domains, neglecting even bigger risks inherent in developing extremely large, capable, agentic, and potentially misaligned AI systems themselves.

The suggestions laid out in

- *“Towards best practices in AGI safety and governance: A survey of expert opinion”* https://arxiv.org/abs/2305.07153
- *“Frontier AI Regulation: Managing Emerging Risks to Public Safety”* https://arxiv.org/abs/2307.03718,
- *“Policymaking in the Pause - What can policymakers do now to combat risks from advanced AI systems?”* https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf

all provide excellent templates from which to approach AI policymaking. They should be read and applied liberally by Australian policymakers.

These suggestions include:

1. Expand technical AI safety research funding (mechanistic interpretability, scalable oversight, etc.)
2. Mandate robust third-party auditing and certification on the safety of the largest frontier-pushing AI systems, using independent dangerous capabilities evaluations
3. Regulate computational power used in training the largest frontier-pushing AI systems
4. Establish liability for AI-caused harms

Target areas

9. Given the importance of transparency across the AI lifecycle, please share your thoughts on:
a. where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?
b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

While transparency is important in the organisations developing and deploying AI systems, it is crucial (and crucially lacking) in the AI systems themselves. If the outputs of an AI system are not correctly explainable, then no amount of institutional transparency will yield any practical transparency or warrant public trust. Crucially, the explanations of an AI’s outputs or behaviour must be sufficiently mechanistic and predictive, rather than post-hoc just-so stories. Sufficiently rigorous explanations are useful not only for those immediately affected by decision-making AI systems, but also for society at large to be sure that AI systems are behaving as they do for the right reasons, so we can have confidence that they will generalise correctly to new situations in high stakes. However it is surprisingly easy to have the “illusion of interpretability” when explaining the behaviour of large AI systems, even when analysing the internal processes going on within them, as shown by Bolukbasi et al. (https://arxiv.org/abs/2104.07143). Explanations must be:

- *Scientific* - sufficient attempts must have been made to falsify the explanation, and all reasonable alternative explanations must have been falsified.
- *Predictive* - The explanation should be sufficient to predict the relevant behaviour again in related new situations.

Sufficiently satisfying these criteria for advanced systems will require much more research in mechanistic interpretability or related fields, which the Australian government should fund.

10. Do you have suggestions for:
a. Whether any high-risk AI applications or technologies should be banned completely?
b. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

A ban should be imposed on the training of any large-scale AI systems with a nontrivial expected chance of passing capabilities evaluations required to cause catastrophic risks to civilization (such as the Alignment Research Center's tests for the capabilities required for autonomous replication https://evals.alignment.org/), until a consensus is reached among alignment researchers that AI systems at this level of capabilities can be made safe.

This ban would only affect the extremely large training runs pushing the frontiers of foundation models, and would only come into effect when the chances of passing dangerous capabilities evaluations are independently deemed nontrivial (though this may be quite soon). For increased safety and simplicity, an immediate ban could be placed on the use of more than a set amount of compute (for example 10^{25} FLOP) in the training of any AI system. This "compute ceiling" could be raised over time as a consensus is formed around technical alignment solutions at that scale, or lowered over time as algorithmic efficiencies allow more dangerous capabilities to arise at lower compute costs.

All current AI work in Australia would be completely unaffected by such a ban, but it would prevent Australia from becoming a harbour for the training of AI systems which pose catastrophic risks to humanity, and set a precedent for other countries around the world.

11. What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

The best way to be trusted is to be trustworthy. When many of the people closest to the cutting edge of AI development and research are sounding the alarm on existential risks from AI, it is not unreasonable for the public to demand more action before they trust that rapid development and deployment of more advanced AI systems will result in a positive future.

That said, a distinction should be drawn between the largest AI systems which push the cutting edge and may soon be extremely dangerous, and relatively mundane small application-based AI systems which form a majority of the current Australian-developed AI systems.

Implications and infrastructure

12. How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia's tech sector and our trade and exports with other Countries?

This question misunderstands both the sources of risk and the interests of Australians.

There are two potential sources of risk both of which need to be mitigated. The first source of risk, as identified in the question, is the use case. Some uses for AI are risky and should be restricted or banned. The second source of risk is the technology itself. Some kinds of technology are inherently risky, and merely restricting a use case does not adequately mitigate the risk. Instead, Australia should insist on dangerous capabilities evaluations of all large AI systems during training, aided by mechanistic interpretability tools, so that we can be aware when we are approaching this extremely risky period and react accordingly.

The economic framing also misunderstands Australia's national interests. We would never conceive of a conversation about whether our aviation sector would trade more successfully if we authorised airlines to use planes that are untested, experimental or known to be dangerous (especially if everyone on earth is potentially on the same plane, metaphorically speaking). Australia's national interest is best served by being on the front foot in how we regulate, being leaders in AI Safety, and shaping the global conversation to favour transparency.

Risk-based approaches

14. Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?

Yes, I support a risk-based approach, but the existential risks I am worried about originate when sufficiently large and capable AI systems are trained, not just when they are deployed. Therefore, for sufficiently large systems with the potential for catastrophic capabilities, the risk-based approach must also be applied before the training phase - analysing risks depending on the nature and scale of the proposed training run, rather than merely analysing the potentially risky use-cases and deployment issues of an AI after it has already been created and trained. Australia should insist on dangerous capabilities evaluations of all large AI systems to be deployed here, but these evaluations must occur during training, aided by mechanistic interpretability tools, so that we can be aware if AI systems are becoming extremely risky and react accordingly

Any assessment of risk should consider its potential irreversibility, scope and severity, and act proportionally.

15. What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

Risks need to be understood before something goes wrong, not after. A potential "worse-case scenario" is if a culture of "trial and error" is adopted, whereby adverse events occur before risks are considered tangible enough to act upon. It is not enough to only react once the risk has been actualised - for some risks that's too late.

Risks also need to be defined sufficiently broadly to encourage caution, and the reaction to a risk should be proportional to the potential level of harm. It may also be wise to include "hedges" against scenarios that are unlikely or uncertain – but highly consequential. That is, if we cannot be sure if a scenario is likely or not, but we know it would be catastrophic, including mitigations to it in our "portfolio" is prudent.

Proliferation of risks is a key point in governing AI, because regulating software poses significant challenges compared to other dual-use commodities such as pharmaceuticals or firearms. This means that, from the outset, there must be proactive regulatory oversight for the deployment of advanced AI systems and their precursors.

Leading researchers from Google and OpenAI have advised that the level of caution required in deploying advanced AI systems should be similar to "caution observed for nuclear power plants, military aircraft carriers, air traffic control, and other high-risk systems" [1].

The report (provided below), provides an excellent introduction to the problems that AI safety researchers face in ensuring that advanced AI is safe, reliable, and includes robust guardrails against misuse. It would be highly beneficial for the Australian Government to take this research into account when developing an understanding of risks that need to be mitigated.

1. Hendrycks et al. Unsolved Problems in ML Safety (2022): https://arxiv.org/pdf/2109.13916.pdf

16. Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?

As the scale and capabilities of AI systems increase, a risk-based approach should switch from mainly focusing on deployment risks in narrow domains to also considering risks inherent in the development of the AI systems themselves. It is important that a true risk-based approach is implemented, rather than an approach blind to these most impactful risks.

The Government should require that developers of the most advanced large-scale AI systems comply with a stringent pre-development review process in order to legally distribute their models in Australia. This process should begin before development has commenced, and later confirming that the organisation has undergone the necessary assurance processes (e.g. forecasts of potential dangerous capabilities before development, experimental dangerous capabilities assessments during development, third-party audits, explainability, etc). This may require a new Australian AI regulator to oversee compliance, and to set the boundaries of which AI systems require this level of precaution (potentially by thresholds on the amount of compute used during training).

There is a large body of evidence that should inform the design of this regulatory process. Examples include:

- This review of current challenges that developers of advanced AI will not to navigate in order to safely deploy their systems: Hendrycks et al. (2022); Unsolved Problems in ML Safety https://arxiv.org/pdf/2109.13916.pdf

- This survey of expert opinion on best practices in AI governance: Schuett et al. (2023); Towards best practices in AGI safety and governance https://www.governance.ai/research-paper/towards-best-practices-in-agi-safety-and-governance

It will also be important to work with organisations in the compute supply chain to help ensure that responsible actors have access to compute resources, while malicious or negligent actors are prevented from having access to undertake dangerous activities, such as training “black market” advanced systems, or using them for malicious purposes.

The following paper provides a valuable overview of compute governance:
Shavit, Y (2023): Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring; Available: https://arxiv.org/pdf/2303.11341.pdf

17. What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

When the potential risks are extremely high (eg. comparable with nuclear war), risk assessments should be carried out prior to (and during) the training of the AI system, and should be reviewed by many more than a few peers. Mandatory third-party auditing for these extremely large and potentially catastrophic training runs is vital.

As for the elements presented:
“Human in the loop” is extremely unlikely to be a sufficient or useful intervention to prevent risks at this scale, especially when capability thresholds for deceiving humans are passed.

“Decision making” may also be the wrong way to frame these risks, unless this phrase is interpreted very broadly. Frontier systems seem likely to do much more than just aid decision making, unless their development or deployment is restricted. “Behaviour” could be a more-general term encompassing not only decision making but all types of outputs from AI systems (including text or image generation, actions in virtual or physical environments, etc.).

“Explanations” however is an excellent requirement, as long as the explanations are required to be sufficiently mechanistic and predictive, rather than post-hoc. Sufficiently rigorous explanations are useful not only for those immediately affected by decision-making AI systems, but also for society at large to be sure that AI systems are behaving as they do for the right reasons, so we can have confidence that they will generalise correctly to new situations in high stakes. However it is surprisingly easy to have the “illusion of interpretability” when explaining the behaviour of large AI systems, even when analysing the internal processes going on within them, as shown by Bolukbasi et al. (https://arxiv.org/abs/2104.07143). Explanations must be:

- *Scientific* - sufficient attempts must have been made to falsify the explanation, and all reasonable alternative explanations must have been falsified.
- *Predictive* - The explanation should be sufficient to predict the relevant AI outputs or behaviour again in related new situations.

Sufficiently satisfying these criteria for advanced systems will require much more research in mechanistic interpretability or related fields, which the Australian government should fund.

18. How can an AI risk-based approach be incorporated into existing assessment frameworks (like privacy) or risk management processes to streamline and reduce potential duplication?

For extremely catastrophic risks from AI, the closest existing processes and policies (such as those for preventing nuclear war) are quite distant. New approaches, policies, or agencies should be created to deal specifically with these risks. Certainly though, no AI-specific regulation should ever waive existing safeguards. Rather, there are entirely new emerging risks which must be dealt with additionally with new regulation, or the formation of a new regulating body.

19. How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

The risk-based approach must be modified to address risks from general purpose AI systems, especially LLMs and MFMs. These types of systems learn their capabilities and behaviours during training, which can be unforeseen by developers. That means that AI developers could create models with dangerous capabilities (e.g., deception, manipulation, political strategy, weapons design). Humans could misuse those capabilities to cause harm [1], and models could apply the capabilities even without deliberate misuse (e.g., through misalignment) [2].

Consistent with recent work from a consortium of AI safety researchers, including at Google Deepmind, OpenAI, Anthropic, Alignment Research Center, University of Cambridge, University of Oxford, Centre for Long-Term Resilience, and Centre for the Governance of AI [1], any risk-based approach should require that general purpose models are evaluated for dangerous capabilities and misalignment. This will help determine:

1. To what extent a model is capable of causing extreme harm (which relies on evaluating for certain dangerous capabilities).

2. To what extent a model has the propensity to cause extreme harm (which relies on alignment evaluations).

The results of these evaluations could be used to determine the level of risk involved in the development and use of these models.

First, evaluations could be used to determine actions that must be taken for responsible training of capable general purpose models, such as through pausing or delaying training, or adjusting training methods (algorithms, data, alignment techniques).

Second, evaluations could be used to inform a deployment risk assessment, and therefore activate certain guardrails before, during, and after deployment (e.g., if updating the model, or by monitoring the model behaviour for examples of misuse or misalignment).

Third, evaluations can also be used to support transparency measures such as requiring incident reporting for higher-risk general-purpose models, sharing pre-deployment risk assessment, scientific reporting, and/or educational demonstrations.

Finally, models at risk of exhibiting dangerous capabilities require strong and novel security controls. Examples of these security controls include red teaming, monitoring of system behaviour and use, network isolation, rapid response to model actions, and verification of system integrity.

Overall, the Government should require that developers of the most advanced large-scale AI systems comply with a stringent pre-development review process in order to legally distribute their models in Australia, regardless of whether the development took place in Australia. This process should begin before development has commenced, including e.g. forecasts of potential dangerous capabilities before development, experimental dangerous capabilities assessments during development, third-party audits, sufficient explainability tools, etc. This may require a new Australian AI regulator to oversee compliance, and to set the boundaries of which AI systems require this level of precaution (potentially by thresholds on the amount of compute used during training).

This calibrated, risk-based approach allows innovation while managing unprecedented risks. It enables beneficial use of AI where appropriate, while restricting high-risk applications pending safety assurances. With advanced AI on the horizon, it is imperative to act preemptively to protect the public.

[1] Brundage et al (2018). “The malicious use of artificial intelligence: Forecasting, prevention, and mitigation”. https://arxiv.org/abs/1802.07228

[2] Ngo et al (2023). “The alignment problem from a deep learning perspective.” https://arxiv.org/abs/2209.00626

[3] Shevlane et al (2023). “Model evaluation for extreme risks”. https://arxiv.org/abs/2305.15324

20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to:
a. public or private organisations or both?
b. developers or deployers or both?

At least for the largest AI systems pushing the frontiers of AI capabilities, a risk-based approach which takes into account the highest-impact risks (unexpected dangerous capabilities, goal-misalignment) should be mandated through regulation primarily focused on developers rather than deployers, even if those developers are overseas.

This is a collective action problem: by gambling with or ignoring these catastrophic risks, individual AI developers will be able to further their own short-term interests. A scheme that is voluntary or driven by aspirational best practices is unlikely to be successful. Businesses may be pressured by market forces to develop or deploy AI tools in risky ways that they don’t necessarily understand to remain competitive with the market. Overall, non-regulatory approaches need to operate hand-in-hand with a strong regulator. And regulation needs to reach back to developers, and not just target deployers and users.

In general, regulation should:

1. Be proportionate to risk. Meaning that a voluntary or self-regulation approach should only apply to low-risk or no-risk uses of AI.

2. In the case of higher-risk technologies, regulate both the technology itself and the use of the technology. Meaning that, for future and more sophisticated AIs, regulation should ensure that the technology is safe before it is created or published, as well as ensuring that it is only used in ways that are safe.

3. Place burdens on those most able to reduce potential harms. Meaning that AI that functionally operates as ‘black box’ to consumers has to be regulated at the point of the AI Lab.

Applying those three principles, self-regulation may be appropriate for many technologies available today and for many participants (such as students, researchers, or AI applications companies not pushing into new AI capabilities from the largest, least well understood AI systems), but a strong forward-looking regulatory regime must apply otherwise, even to AI systems developed outside of Australia.