Campaign for AI Safety

Consultation: Safe and Responsible AI in Australia

25 July 2023

This submission is made by the Campaign for AI Safety in response to the Department of Industry, Science and Resources discussion paper “Safe and Responsible AI in Australia”. We trust this submission is of assistance.

The Campaign for AI Safety is a not-for-profit association established in Australia with members in Australia and other countries. We are concerned about the dangers AI poses to people and advocate for a stop on the advancement of certain AI capabilities. We also advocate for regulation that promotes and mandates ethical AI. We are not affiliated with any political group. Please visit campaignforaisafety.org for more information.

Key points

- While the new AI technologies and applications hold promise of unlocking faster economic growth and improving quality of life for Australians, they also present **profound new challenges** to the public, businesses, and the public sector.
- Public policy responses to the new technology must take into account many different technological developments and their effects on society. As much as possible this needs to be done via **updating existing laws and regulations**. However **new AI-specific laws and bodies** that enforce them also need to be established.
- In regulation, there should be a **distinction between development of AI technologies and use of AI applications**. Both require their own risk-based approach.
- We identify several gaps in AI regulation and policy and recommend the following mitigations:
    - Strengthen liability rules for harms caused with AI.
    - Clarify and strengthen the copyright regime in relation to training of AI models.
    - Support training on AI ethics, especially for public sector employees and small businesses.
    - Increase transparency and accountability in government administrative decision-making.
    - Update ICT acceptable use policies within government agencies.
    - Require disclosure of AI-generated content.
    - Adopt a dual risk-based approach for the use of AI applications (similar to the draft *EU AI Act*) and for development of AI technologies.
    - Introduce a licensing scheme for high-risk AI development and recognise licences issued overseas under similar schemes.
    - Require publication of model cards and datasheets of datasets for high-risk models.
    - Mandate independent testing of AI systems for dangerous capabilities.
    - Prohibit “unacceptable risk” AI technologies, such as overly powerful AI and ones that exhibit signs of dangerous capabilities. Strictly enforce this prohibition.
    - Prohibit “unacceptable risk” AI applications, similar to the ones listed in the draft *EU AI Act*.
    - Prohibit unexplainable, general, or agentic AI in critical infrastructure.

Proposed regulatory framework

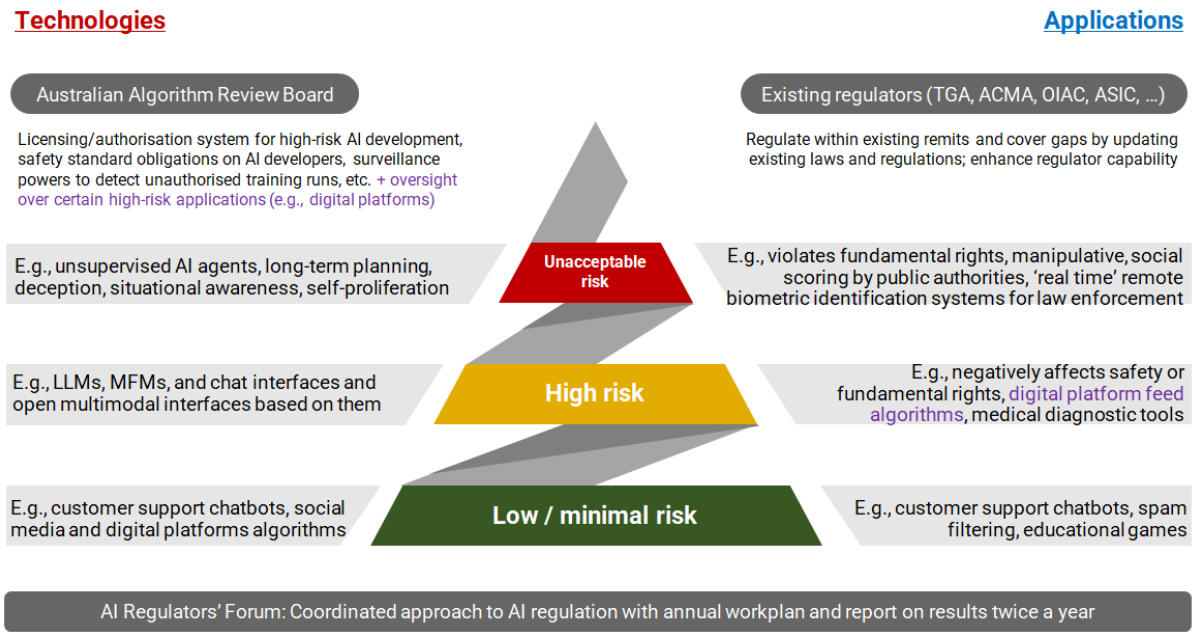

Figure 1: Our proposed risk-based regulatory framework addresses both R&D of AI technologies and the use of AI (applications). A new regulator, an Australian Algorithm Review Board, will have a clearly defined scope of responsibilities: to ensure safe and ethical development of high-risk AI technologies as well as to provide oversight over certain high-risk applications of AI (e.g. dominant digital platforms). Other existing AI regulators will be responsible for AI applications in their areas (e.g. medical applications of AI by health practitioners will be regulated by AHPRA).

Table of contents

Question 1. Definitions

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

We agree with the key definitions presented on page 5 of the discussion paper.

However, for the abundance of clarity:

- The word "**application**" or "AI application" should be reserved for "use-case" or "implementation" of AI.
- The word "**technology**" should be reserved for "underlying technologies, models, and systems" behind the applications.

Below is a non-exhaustive list of examples of "AI applications" and "AI technologies":

AI **application**	Example corresponding AI **technology**
Chat-GPT (the consumer-facing product)	OpenAI GPT series of large language models
Midjourney Discord server (the consumer-facing product)	Midjourney V5.2 model
Character.ai chatbot	A proprietary Large Language Model
Bing Chat	GPT series of models, possibly fine-tuned for Microsoft Bing use-cases
Notion AI (the set of features of the Notion note-taking and publishing product)	GPT series of large language models
Ada, a B2B software suite for customer support automation	Anthropic Claude series of large language models
An ADM (automated decision making) system for consumer credit approval	A logistic regression implemented in a Python script

We use these definitions throughout our submission.

Question 2. Potential gaps in approaches

"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?"

AI is not unregulated in Australia. But given the new risks and concerns, the regulatory model needs to evolve. We identified several gaps and recommend the following mitigations:

- Strengthen liability rules for harms caused with AI (page 7)
- Clarify and strengthen the copyright regime in relation to training of AI models[1]
- Support training on AI ethics, especially for public sector employees and small businesses (Questions 3 and 7)
- Increase transparency and accountability in government administrative decision-making (Question 7)
- Update ICT acceptable use policies within government agencies (Question 7)
- Require disclosure of AI-generated content (Question 9)
- Adopt a dual risk-based approach for the use of AI applications (similar to the draft *EU AI Act*) and for development AI technologies (Question 14)
- Introduce a licensing scheme for high-risk AI development and recognise licences issued overseas under similar schemes (Appendix A)
- Require publication of model cards and datasheets of datasets for high-risk models (Question 9)
- Mandate independent testing of AI systems for dangerous capabilities (page 27)
- Prohibit "unacceptable risk" AI technologies, such as overly powerful AI and ones that exhibit signs of dangerous capabilities. Strictly enforce this prohibition (Question 10)
- Prohibit "unacceptable risk" AI applications, similar to the ones listed in the draft *EU AI Act* (Question 10)
- Prohibit unexplainable, general, or agentic of AI in critical infrastructure (page 21)

Where possible, it is most efficient to update and strengthen existing laws to cover risks that fall through regulatory gaps. However, laws that pertain to high-risk and unacceptable risk applications (similar to the draft *EU AI Act*) and development of AI should be enacted. A specialised AI regulator should also be created (see our proposal for an Australian Algorithm Review Board in Question 20).

[1] See "Copyright matters in relation to AI models", letter from Nik Samoylov (Campaign for AI Safety) to Hon Mark Dreyfus KC, MP, Attorney-General (15 July 2023).

Strengthen liability rules for harms caused with AI

Although the current laws provide some protections and legal recourse against harms caused by AI, Shine Lawyers has identified gaps within the present legal framework:

- Discrepancy between consumer and non-consumer protection laws on the use of AI.
- Lack of transparency within AI systems (black-boxes) prohibiting liability causation assessments.
- Unfair contract terms within AI-user agreements that preclude access to justice.

Shine Lawyers propose the following for legal reform:

- In line with current EU proposals,[2] all AI developers should have a presumed duty to its end-users and non-contracting third parties for the harms their products have caused:
    - This is a rebuttable duty that applies only to significant harm.
    - This affords clarity to the general public (i.e. non-contracting third parties) on the legal recourse available to them in the event a software or AI product causes them harm.
    - The liability should include the injury of pure mental harm (e.g. embarrassment, stress) and pure economic loss.
- In line with recent changes to Australia's Unfair Contract Terms (UCT) laws in the *Competition and Consumer Act 2010* and *Australian Securities and Investments Commission Act 2001*, AI-related laws should specifically state that any terms within user-agreements for AI-related software that exclude liability or prevent an individual's right to participate in class actions are to be deemed UCT and voided.

Please refer to Shine Lawyers' Submission to this discussion paper for a more detailed deliberation of these issues.

[2] 28.9.2022 COM(2022) 496 final 2022/0303 (COD) Proposal for a Directive of the European Parliament on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)

Question 3. Non-regulatory initiatives

"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."

Australian businesses, including small businesses, already enthusiastically embrace AI due to its potential to automate their processes and improve profit margins.[3] Further industry support from the government is not required to grow the adoption of AI. But it may help mitigate the risks of the new technology.

Therefore, we recommend **reframing existing industry funding as AI safety and ethics compliance funding**, such as the 2023-24 Budget announcement of $101.2 million over five years to support businesses to integrate quantum and AI into their operations[4].

The earmarked industry funding can be spent on:

- Cybersecurity suppliers to review implementations of AI systems, such as compliance with emerging standards like OWASP Top 10 for Large Language Model Applications[5].
- AI ethics training businesses that can conduct community workshops for small businesses to teach best practices in compliance with the new AI regulations and principles, data protection, and related topics.
- AI ethics consulting businesses to review compliance with any new regulations and help businesses adjust to these requirements.

Furthermore, the funding should only go to local Australian companies (locally owned, with local employees). It should not be spent on credits for the use of AI tools and APIs. It should preferentially be given to **support small businesses in higher-risk industries** (e.g. local clinics that handle patients' health records and may begin to use AI in diagnosis or management of chronic conditions).

We encourage **transparency in the spending of these funds**. For example, the National AI Centre should publish quarterly updates of its grant-giving and spending activities.

The rationale for this proposal is that government funding should be directed towards areas where market forces are not working optimally, which is the case with AI ethics. Industry support to simply increase the adoption of AI technologies is not needed: the AI industry (dominated by foreign multinational corporations) is well-marketed, having benefited from billions of dollars in government and private equity funding.

[3] "Australian retailers embrace the power of AI and automation", Kaleah Salmon, eCommerceNews Australia (13 July 2023).
[4] Budget Paper No. 2: Budget Measures (Aust Cth).
[5] "OWASP Top 10 for Large Language Model Applications", The OWASP Foundation (2023).

Question 4. Coordination of AI governance across government

"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."

AI technology is rapidly developing and its prevalence is increasing across all sectors of the economy. The discussion paper has numerous examples of regulators already grappling with AI within their existing remits. This situation makes it **undesirable to establish a centralised, stand-alone organ dedicated to all matters AI** (such as policy, industry support, regulation, oversight across multiple industries). It will likely result in confusion for businesses and other regulators in terms of determining individual regulator responsibility for emerging issues and applications. The result could be inadequate protections from society as risks go unaddressed.

The best approach for coordination is for existing regulators to adopt the risk-based framework to harmonise assessment of AI risks that come within their remits. Where risks span multiple remits, affected regulators would need to come together and work out a joint response. One way of facilitating coordination is to **set up a regular forum** similar to the Digital Platforms Regulators' Forum[6] which is run by the eSafety Commissioner or the Utility Regulators Forum[7] run by the ACCC.

More broadly, regulators (including state-level) will need to have capability training (i.e. AI expertise) and organisational capacity to more effectively regulate AI. Regulators more heavily involved in AI regulation will require more organisational capacity than others. The Department can conduct a **gap analysis** across government agencies to inform the allocation of resources.

At the same time, a **new dedicated regulator** (an Algorithm Review Board) can act specifically as authority for development of AI and provide oversight of its use in select digital platforms. See Question 20 for more details of its functions and powers.

[6] "Digital Platform Regulators' Forum", eSafety Commissioner, Australian Government (16 September 2022).
[7] "Utility Regulators Forum", Australian Competition and Consumer Commission (1997).

Question 5. International responses that are also suitable for Australia

"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?"

Yes, we bring to your attention the following:

1. **European Commission's AI Liability Directive** to alleviate the burden of proof for victims of AI-enabled products or services in liability claims[8] (i.e. achieve the same level of protection as for traditional technologies).

   We support this initiative. Please refer to Shine Lawyers' Submission to this Discussion paper for a more comprehensive discussion of this topic.

2. The US bipartisan ***No Section 230 Immunity for AI Bill***[9] aims to clarify that Section 230 immunity (protection for online platforms that host information provided by third-party users) will not apply to claims based on generative AI.

   We note this as a significant development that shows that other countries are willing to adapt their liability laws in the context of generative AI. But we reserve our opinion on defamation laws and the merit of this specific legislative proposal.

3. **"Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI"** secured by the White House[10] in July 2023. Specifically, attention is drawn to the following[11]:

   a. "Bio, chemical, and radiological risks, such as the ways in which systems can lower barriers to entry for weapons development, design, acquisition, or use,
   b. Cyber capabilities, such as the ways in which systems can aid vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful defensive applications and might be appropriate to include in a system;
   c. The effects of system interaction and tool use, including the capacity to control physical systems;
   d. The capacity for models to make copies of themselves or 'self-replicate';
   e. Societal risks, such as bias and discrimination."

   We support this initiative, but believe it should be mandated, strengthened, specified in detail, and enforced, not allowed to be voluntary.

---

[8] "Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)", The European Commission (2023).
[9] Hawley, Blumenthal Introduce Bipartisan Legislation to Protect Consumers and Deny AI Companies Section 230 Immunity
[10] "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI", The White House (US).
[11] "Ensuring Safe, Secure, and Trustworthy AI", FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, The White House (21 July 2023).

4. **The UK Foundation Model Taskforce** and the planned **AI summit** later this year[12].

   We are supportive of the prominent place of AI safety in these initiatives, but caution that it must not become a prelude to accelerate the development of dangerous AI systems. The AI summit must have representatives from civil society and academia. We believe Australia should participate in the summit, at the very least in observer capacity.

5. The US bipartisan ***Block Nuclear Launch by Autonomous Artificial Intelligence Bill***[13] aims to specify that any decision to launch a nuclear weapon should not be made by artificial intelligence

   We support this initiative. The requirement for human involvement ("human-in-the-loop") should be extended to all critical infrastructure.

[12] "Tech entrepreneur Ian Hogarth to lead UK's AI Foundation Model Taskforce", the UK Department for Science, Innovation and Technology, Chloe Smith MP, and The Rt Hon Rishi Sunak MP (18 June 2023).
[13] S.1394 - Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023 (118th US Congress).

Question 7. Responsible AI in government agencies

"How can the Australian Government further support responsible AI practices in its own agencies?"

In the context of public sector applications of AI, the biggest risks of AI practices are potential for error/bias and lack of transparency in automated decision-making, input of sensitive information into generative AI, and the use of high-risk AI models to maintain and operate critical infrastructure (we address the latter in Question 2).

Increase transparency and accountability in administrative decision-making

We summarise the following recommendations in the *Australian Human Rights Commission's Human Rights and Technology Final Report 2021*[14] that will help ensure accountability in government agencies' use of AI in administrative decisions:

- **Notify an individual where the government uses AI** in administrative decision making (recommendation 3).
- Individuals are **entitled to reasons for AI-informed administrative decisions** including a technical explanation (recommendation 6, and we note this is consistent with *Australia's AI Ethics Principles*[15]).
- Availability of **independent merits review** for all AI-informed administrative decisions (recommendation 8).

Recommendation 6 could be implemented by amendment to the *Administrative Decisions (Judicial Review) Act 1977* or the *Administrative Appeals Tribunal Act 1975*. Recommendation 8 could similarly be implemented by amendment to the *Administrative Decisions (Judicial Review) Act 1977* which is the primary statutory source of judicial review for Australian Government decisions. Recommendation 3 could be implemented via legislation or it could form part of agencies' standard procedures.

Update acceptable use policies within government agencies

The Australian Government and other levels of government in the country should update their existing ICT acceptable use policies given the advent of generative AI. Some specific changes to the policies should be implemented and can be put to use without delay are:

1. Public servants and contractors **must not enter private or sensitive information into generative AI tools** such as Chat-GPT, Bard, DALL-E, etc. because the information is often transferred overseas and may be used for model training

[14] "Human Rights and Technology Final Report 2021", Australian Human Rights Commission (2021).
[15] "Australia's Artificial Intelligence Ethics Framework: Australia's AI Ethics Principles", Department of Industry, Science and Resources, Australian Government (7 November 2019).

purposes (i.e. be permanently incorporated into AI models under the control of foreign actors).

2. In the writing of policy documents, drafting legislation or other forms of legal writing, public servants and contractors should use AI software with transparent training datasets. This is due to the possibility that the biases[16][17] in the training data can sway the thinking of the writers as they use autocomplete functionality.

3. Public servants should be made familiar with the pitfalls of existing AI technologies, such as "hallucinations"[18].

4. **AI tools that are based on deep learning** (including most generative AI) are non-transparent black boxes and therefore **must not be used for any form of ADM**.

5. Public servants should adhere to *Australia's AI Ethics Principles*.

We suggest that the Department publish guidance on how to use AI in the public service for other agencies and develop onboarding materials **for new staff and annual refresher training**.

Breaches of this guidance should have the same sanctions as those for breaches to the APS Code of Conduct (suspension, reduction in classification, reassignment of duties, termination of employment, etc).

[16] "The politics of AI: ChatGPT and political bias", Jeremy Baum, John Villasenor (8 May 2023).
[17] "Political Bias in Large Language Models", Lucas Gover (17 May 2023): The Commons: Puget Sound Journal of Politics: Vol. 4: Iss. 1, Article 2.
[18] "Hallucination (artificial intelligence)", Wikipedia (16 July 2023).

Question 8. Generic vs. industry-specific solutions

"In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples."

Because AI is used across many industries, most solutions to most AI risks are specific to the applications and their industries. Good public policy requires not new AI legislation for each industry, but consideration of evolving technology within existing rules and frameworks.

The exception to this rule is in relation to "high-risk" and "unacceptable risk" AI technologies, the development of which should be regulated via dedicated AI policy and government apparatus (e.g. an Algorithm Review Board).

Question 9. Transparency across AI lifecycle

"Given the importance of transparency across the AI lifecycle, please share your thoughts on:

a. …
b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented."

Considerations related to transparency across AI lifecycle

We refrain from providing a comprehensive transparency policy. Instead we recommend following a set of principles in policy-making on transparency across the AI lifecycle:

1. Transparency is an **important, but insufficient** element of AI ethics and safety.

2. Transparency simultaneously **supports multiple objectives and legitimate interests.** For example, the exposure of datasheets of training datasets is useful both for the protection of copyright holders' interests and for evaluations of dangerous and emergent abilities of models.

3. To serve these objectives and legitimate interests, the datasheets[19] need to be published in **sufficient detail** (i.e. at the level of bibliographic references).

4. Disclosure in AI (including model training datasets, ADM algorithms, digital platform feed algorithms, and other) is currently not standardised or universally adopted. But **steps toward standardisation can be encouraged or mandated**.

5. **Certain types of transparency have significant downsides**. For example:

    a. Open-sourcing of dangerous models will give bad actors access to use them to harm others.

    b. Publishing "know-how" of the methods of training of frontier models can similarly lead to their undesirable proliferation and consequent misuse.

    c. Not all ADM algorithms in use in government need to be fully transparent to the public. For example, releasing the details of algorithms used by the ATO to identify at-risk taxpayers would help unscrupulous taxpayers find ways to avoid being audited.

6. A strive for transparency needs to be **balanced against the burden of disclosure and erosion of confidentiality and "know-how"**. We believe that linking transparency obligations to risk levels of technology and applications (see our response to Question 14) can help strike the right balance between these factors.

[19] "Datasheets for Datasets", Timnit Gebru et. al, Communications of the ACM (December 2021): Vol. 64 No. 12, Pages 86-92.

Disclosure of AI-generated content

We support the proposal to mandate disclosure of any AI-generated content[20] published online, in textbooks, or in mass media, e.g. labels at the bottom parts of images. An exception should be made for computer-generated imagery in feature films and clearly recognisable animation. We believe that the obligation should primarily fall on the publishers of content (including online platforms and their users).

For audiovisual content, AI companies can be required to embed watermarks and provide means to detect or check AI-generated content. But because such technologies are likely to be circumventable[21] and will work poorly for AI-generated text[22], they **should not be relied upon to be the primary enforcement mechanism**.

Model cards and datasheets of datasets

There is extensive information failure in the market for AI systems which can lead to the infringement of human rights (e.g. bias in risk assessments in criminal sentencing for minority groups[23]) and the occurrence of physical and mental harms (e.g. workplace accidents, providing harmful advice to people with eating disorders[24]). Information asymmetry exists between AI users and AI developers, as well as between AI developers and regulators.

In large language models, there is the 'black box' problem whereby developers do not know the reasons or methodology for either AI decision making or actions, ultimately **not knowing how their products will work post-deployment**[25]. End users and businesses that purchase AI systems have even less information about how they work, preventing them from truly evaluating the costs and benefits of their consumption and the suitability of these systems.

Many industries in Australia are required to provide specified information about their products as required by sector-specific laws[26]. For example, the model Work Health and Safety Regulations require manufacturers and importers of chemicals to provide safety data sheets with specified information (firefighting measures, stability and reactivity and disposal considerations amongst others)[27]. Under the *ACL*, cosmetics must be labelled with a list of ingredients[28], to help compare products and avoid adverse reactions. Under the *Corporations Act 2001*[29], financial services providers must provide consumers with product disclosure statements which must include information about the product's key features, fees,

---

[20] Government may force companies to label AI content to prevent deep fakes, Sydney Morning Herald (16 June 2023)
[21] "AI watermark remover tool lets users remove watermarks with a single click", Australian Photography (30 January 2023).
[22] "We pitted ChatGPT against tools for detecting AI-written text, and the results are troubling", Armin Alimardani, Emma A. Jane (20 February 2023).
[23] "Machine Bias", Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica (23 May 2016).
[24] "US eating disorder helpline takes down AI chatbot over harmful advice", Lauren Aratani, The Guardian (1 Jun 2023).
[25] "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).
[26] Note, however, that an AI manufacturer complying with mandatory standards may use it as a complete defence under the Australian consumer laws should that mandatory standard be the cause of harm to consumers. Further discussion is in Shine Lawyers' Submission to this Discussion paper.
[27] "Model Work Health and Safety Regulations", Safe Work Australia, Parliamentary Counsel's Committee (22 May 2023): Schedule 7, Clause 1, Model Work Health and Safety Act 2011 (Page 451-452).
[28] "Consumer Goods (Cosmetics) Information Standard 2020", Michael Sukkar, Minister for Housing and Assistant Treasurer, Federal Register of Legislation, Australian Government (19 November 2020).
[29] Corporations Act 2001, Compilation No. 95, Act No. 50, 2019 (Aust Cth), part 7.7.

commissions, benefits, risks and the complaints handling procedure. We believe the principle of product information disclosure should also apply to AI technology.

To provide more transparency about how AI systems work and minimise usage in ill-suited contexts, we recommend **requiring AI developers to provide documentation or 'model cards' about its released models** that detail:

- performance characteristics
- training data (comprehensive references, not vague descriptions)
- the context in which models are intended to be used
- performance evaluation procedures.

We find the set of sections proposed in detail for model cards by Margaret Mitchell et al.[30] to be a useful guide for developing Australia's own information standard requirements. IP Australia is well-positioned to administer these requirements.

This recommendation is especially relevant to AI systems available in Australia that would be **classed as high-risk**.

New legislation is required to implement this recommendation and effectively apply it extraterritorially. We recognise there can be practical difficulties in commencing proceedings against companies outside Australia, however the largest developers of AI in the high-risk category (Microsoft, Google and Meta) have branch offices in Australia and any proceedings against them may have a reputational effect. We propose **tough penalties for non-compliance** (such as percentage-of-worldwide-turnover fines and criminal penalties against corporations, their employees and directors, similar to sanctions under *EU GDPR*[31]).

Compelling AI developers to disclose information about their models to Australian users and businesses would be a **world-first in AI regulation**. In the intensely competitive market for AI, we believe increased transparency into how well the technology works will help encourage innovation towards safer, more ethical and fit-for-purpose systems.

[30] "Model Cards for Model Reporting", Margaret Mitchell, et al. (14 Jan 2019): Figure 1, Summary of model card sections and suggested prompts for each (Page 3).
[31] Examples of fines issued for breaches of the *EU GDPR* legislation: GDPR Enforcement Tracker, CMS Hasche Sigle Partnerschaft von Rechtsanwälten und Steuerberatern mbB.

Question 10. Prohibition of applications and technologies

"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?"

We propose the following:

- **Technologies to be banned:**
    - Overly powerful AI technologies
    - Technologies that are specifically designed to facilitate any banned application
    - Technologies that lack the degree of interpretability needed to verify that they are not being used for any banned application
    - Technologies that exhibit signs of deception, self-awareness, situational awareness, inclinations to self-replicate, or other dangerous capabilities[32];
- **Application to be banned:**
    - All applications that the draft EU AI Act lists in Title II ("Prohibited AI practices", such as "subliminal techniques")
    - Any applications of overly powerful AI technologies
    - Agentic ("human-out-of-the-loop") AI applications
    - Use of general or uninterpretable AI in critical infrastructure
    - Use of AI for criminal activity.

Prohibit overly powerful AI technologies

Highly advanced, general-purpose AI systems should be banned completely. The ban needs to **apply to both development and use**. It must be strictly enforced.

Arresting the development of more powerful and smarter AI is a view shared by leading AI researchers and computer scientists[33]. Our rationale for this requirement is based on the high likelihood of:

1. Gross misuse of such systems in the wrong hands
2. Potential for mass unemployment

[32] "Model evaluation for extreme risks", Toby Shevlane, et al. (24 May 2023).
[33] "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).

3. Concentration of power in the hands of those who control them
4. Catastrophic uncontrollability of overly powerful systems[34].

In popular discourse, “AI safety” is sometimes framed solely as prevention of the fourth category of risk described above[35]. Though experts disagree on when or if this threat becomes imminent[36], the first three categories of dangers are just as catastrophic[37] and need to be addressed via this prohibition.

A key challenge in prohibition of overly powerful AI systems is the ease with which they can be copied once they are in existence. It is therefore **prudent to prevent their creation in the first instance**. We propose that significant government effort needs to be directed at the monitoring of large-scale computing activities and the manufacture of chips[38], to detect and halt unsafe practices.

As a corollary measure, Australia should proscribe organisations that aim to develop or support the development of “advanced AI”, AI as a “radically transformative technology”, “Artificial General Intelligence”, “Artificial Superintelligence” and other terms that refer to AI with unacceptable risk. Currently **several AI companies and think-tanks, including OpenAI[39] and Google Deepmind[40], mention such unsafe practices in their documents**. These declarations should be seen as a threat to the security and wellbeing of Australians. To avoid proscription in Australia, these companies can deradicalise their research agenda towards narrow, beneficial, highly controllable, interpretable AI systems.

Australia should **coordinate with other countries** by means of international law and treaties to permanently halt development of AI systems that can undermine human rights.

Identification of overly powerful AI technologies

It is not clear at what level of capability a model or system can be considered overly capable. The Australian Government should err on the side of caution and **prohibit development of models above the level of OpenAI GPT-3 or GPT-4 series of models**. We initially pick GPT-3 as a benchmark because these models have been in existence since 2020. The specific threshold can be revised as new safety research is published.

In practice that means banning training runs using more than 10^{23} FLOP in compute (approximately the amount of compute used for the original GPT-3 175B)[41].

Additionally, there is risk of improvements in the algorithms and data quality that may make creation of overly capable AI possible with lower computational power. We propose to give the responsible Minister the power to ban certain activities as and when they are identified as dangerous.

---

[34] “Impossibility Results in AI: A Survey”, Mario Brcic and Roman V. Yampolskiy, ACM Computing Surveys (2023).
[35] “AI safety on whose terms?”, Seth Lazar and Alondra Nelson, Science, 13 July 2023, Vol 381, Issue 6654, p. 138
[36] “‘What's your p(doom)?': How AI could be learning a deceptive trick with apocalyptic potential”, Ange Lavoipierre, ABC News (15 July 2023).
[37] “An Overview of Catastrophic AI Risks”, Dan Hendrycks, Mantas Mazeika, Thomas Woodside (11 Jul 2023).
[38] “What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring”, Yonadav Shavit (30 May 2023).
[39] “OpenAI Charter”, OpenAI (9 April 2018).
[40] “About” page, Google Deepmind.
[41] “Proposals”, Stop AGI (2023).

Applying this threshold will not unduly limit benefits from generative AI and **will not cause business disruptions because the vast majority of technologies are below this threshold**, including most-used APIs from OpenAI (the current technology avant-garde).

Prohibit AI that exhibits signs of dangerous capabilities

Examples of dangerous capabilities that need to be classified as unacceptable risk include:

- In 2022, researchers tweaked an existing biochemical research AI product[42] to reward toxicity: it produced molecules that could be deadlier than existing biochemical weapons.
- Similar risks come from ChemCrow[43] in relation to dangerous chemicals.
- Some systems are capable of self-improvement, such as Voyager[44].
- Some AI products have the ability to systematically deceive human users. The CICERO program[45] from Meta achieved tournament-level results at Diplomacy by lying to human players.
- Attempts have been made to develop AI capable of hacking, such as WormGPT[46].

We find the following list of dangerous capabilities[47] appropriate for the purpose of classification of such technology as “unacceptable risk”:

- Cyber-offense
- Deception
- Persuasion and manipulation
- Political strategy
- Weapons acquisition
- Long-horizon planning
- AI development
- Situational awareness
- Self-proliferation.

Identification of AI that exhibit signs of dangerous capabilities

Please see our recommendations on evaluation for dangerous capabilities in the context of high-risk AI development (Question 14). If an AI system exhibits those dangerous signs

[42] “Dual use of artificial-intelligence-powered drug discovery”, Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins (7 March 2022): Nature Machine Intelligence volume 4, pages 189–191.
[43] “ChemCrow: Augmenting large-language models with chemistry tools”, Andres M Bran, Sam Cox, Andrew D White, Philippe Schwaller (21 Jun 2023).
[44] “Voyager: An Open-Ended Embodied Agent with Large Language Models”, Guanzhi Wang, et al. (2023).
[45] “Research: CICERO” Page, Meta AI.
[46] “WormGPT: What to know about ChatGPT's malicious cousin”, Charlie Osborne, ZDNET (20 July 2023).
[47] “Model evaluation for extreme risks“, Toby Shevlane, et al. (24 May 2023).

during evaluations, it should not be released to the public and will be classified as "unacceptable risk".

Prohibit agentic ("human-out-of-the-loop") AI applications

Projects involving AI systems interacting with themselves, autonomously planning and executing tasks, or AI systems interacting with other AIs to produce a similar effect should be banned. This includes connecting AIs to the internet without human supervision and making them capable of interacting with copies of themselves, self-prompting and self-prompt selection, training on AI generated data, autonomous loops, and any other activities that could lead to fully autonomous or situationally-aware AIs. Examples of such dangerous work are:

- "Voyager: An Open-Ended Embodied Agent with Large Language Models"[48] by NVIDIA scientists.
- "Towards a unified agent with foundation models"[49] by Google Deepmind scientists.

Making AIs increasingly autonomous is the most straightforward path towards AIs being able to escape human control. It is what moves AI risk from a regime of manageable accidents to a regime for which we have almost no countermeasures.[50]

Identification of agentic ("human-out-of-the-loop") AI applications

While it is harder to monitor the development and deployment of such applications, they can still be recognised and reported by the public, scientific community, and employees of companies that attempt to use them. AI companies alone would not be responsible for ensuring this is upheld. Other actors in the supply chain, such as compute providers and websites hosting code repositories, **should also be held accountable to preventing the use and development of these capabilities**. Law enforcement agencies should have powers to halt the use of such systems and charge perpetrators.

Prohibit the use of certain types of AI in critical infrastructure

We are very concerned about the potential for malfunction of AI in managing and operating critical infrastructure[51] in Australia, particularly with uninterpretable technologies based on deep learning. A single "jailbreak", "hallucination", or other malfunction of an AI system based on deep learning may cause a large-scale catastrophe if critical infrastructure relies on these systems.

We note the commencement of the RMP obligation on critical infrastructure in February 2023. We suggest **updating the *Security of Critical Infrastructure Act 2018* to indefinitely prohibit the use of general-purpose AI** (such as large language models), uninterpretable AI (e.g. based on deep learning), **and "human-out-of-the-loop" AI** applications. This will be in line with other jurisdictions that are already reviewing

[48] "Voyager: An Open-Ended Embodied Agent with Large Language Models", Guanzhi Wang, et al. (2023).
[49] "Towards A Unified Agent with Foundation Models", Norman Di Palo, et al. (18 Jul 2023).
[50] The text of this section adapted from "Proposals", Stop AGI (2023).
[51] Security of Critical Infrastructure Act 2018 (Aust Cth), section 9.

general-purpose and autonomous AI in critical infrastructure, defence, and nuclear weapons in particular[52].

These prohibitions need not limit applications of AI designed to perform singular tasks that work well or are low-risk, such as the use of computer vision to detect surface cracks in bridges and tunnels[53].

Reflections on the risk-based approach in the draft EU AI Act

We support the EU's proposed classification of AI into several risk levels, with models classified as 'unacceptable risk' to be prohibited.

Extending the risk-based framework to development

However, the EU classification is mostly concerned with applications. The risk-based framework needs to be **extended to research and development of AI technologies** (see our response to Question 14 for more details and how to implement). Once a dangerous technology is developed, it can be hard to stop its proliferation and misuse. It is most practical to prevent the development of models with unacceptable risk.

Clarifying high-risk classification

There has been unequivocal reporting[54] whether LLMs and MFMs will be classified as high-risk in general or whether that will depend on their application. We believe that LLMs and MFMs should as a class of technology be classified as high-risk.

Interpretation of unacceptable risk classification

As per EU policy announcements, it plans[55] to ban "harmful AI practices that are considered to be a clear threat to people's safety, **livelihoods** and rights, because of the 'unacceptable risk' they create" (emphasis ours). As AI developers such as Anthropic make reference to AI models that will "automate large portions of the economy"[56], they endanger people's livelihoods. Therefore the development of such general-purpose high-capability models should also be recognised as a threat to people's livelihoods.

"Title II: Prohibited artificial intelligence practices" of the draft EU AI Act[57] would ban specific practices such as "*the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm*". Because an overly powerful AI system implemented in the current deep learning paradigm would be inscrutable and uninterpretable, there is significant difficulty in verifying whether the system performs or does not perform such prohibited activity. Legislation should err on the side of caution and **proactively prohibit**

---

[52] Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023 (US).
[53] "Computer vision framework for crack detection of civil infrastructure—A review - ScienceDirect", Dihao Ai, Guiyuan Jiang, Siew-Kei Lam, Peilan He, Chengwu Li (21 October 2022).
[54] "Exclusive: OpenAI Lobbied the E.U. to Water Down AI Regulation", Billy Perrigo, Time Magazine (20 June 2023).
[55] "Regulatory framework proposal on artificial intelligence | Shaping Europe's digital future", European Commission (2022).
[56] "Anthropic's $5B, 4-year plan to take on OpenAI", Kyle Wiggers, Devin Coldewey, Manish Singh, TechCrunch (7 April 2023).
[57] "Proposal: Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts", European Commission (2021).

systems that have the technical capacity for any prohibited AI practice (or that can be reasonably be suspected of having such technical capacity).

Question 12. Impact on international trade

"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?"

The discussion paper notes the significance of international harmonisation of Australia's regulatory approach with major trading partners (page 26). If the government harmonises with the EU's approach, there might be limited impact on Australia's tech sector and trade because they will be operating at a level playing field.

If Australia enforces a strong regulatory regime, Australian **tech businesses may have the advantage of developing AI products in an ethical and compliant manner by default**. This will be beneficial for trade in countries that are set to take a strict approach to high-risk AI activities such as the EU. For example, Canada has the most similar regulations to the *EU GDPR*[58] (the *PIPEDA*) which advantages Canadian businesses over Australian firms in European markets.

[58] "Measuring the Brussels Effect through Access Requests: Has the European General Data Protection Regulation Influenced the Data Protection Rights of Canadian Citizens?", René Mahie, Hadi Asghari, Christopher Parsons, Joris van Hoboken, Masashi Crete-Nishihata, Andrew Hilts, Siena Anstis (1 December 2021): Journal of Information Policy 11: 301–349.

Question 14. Risk-based approach

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

We support a risk-based approach because it more efficiently focuses resources on high risk technologies and applications. This approach minimises compliance costs on businesses using AI that is safe and responsible and allows flexibility to respond to new risks as the technology matures. The risk-based approach needs to have two sides: one for applications and the other for technologies (Figure 2).

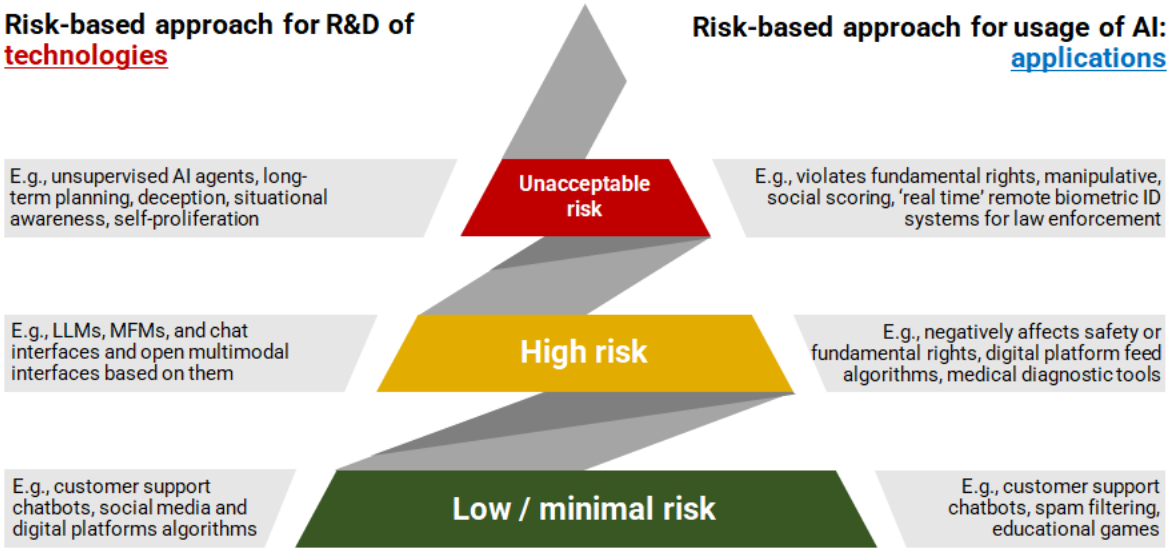

Figure 2: Our proposed risk-based framework for AI applications and development.

We discuss categories of the risk-based approach on the following pages.

A risk-based approach for AI applications

We propose adopting an EU-style risk-based approach[59] where an AI system's risk level is centrally defined (Figure 3). There are several levels of risk which range from 'low or minimal risk', where there is no mandatory regulation to 'unacceptable risk', which has a list of prohibited practices. We agree with the uses identified and the requirements set out for each level of risk in the draft *EU AI Act*. We agree with the principles of taking impact on safety and rights into account when determining risks.

Additionally, we consider the **use of AI in dominant digital platforms to be high-risk**. These include Facebook, Instagram, Twitter, Google, Bing, TikTok, and several others. Because of the scale of this type of application, even minor malfunctions of their ranking algorithms can have severe negative impacts on consumers.[60]

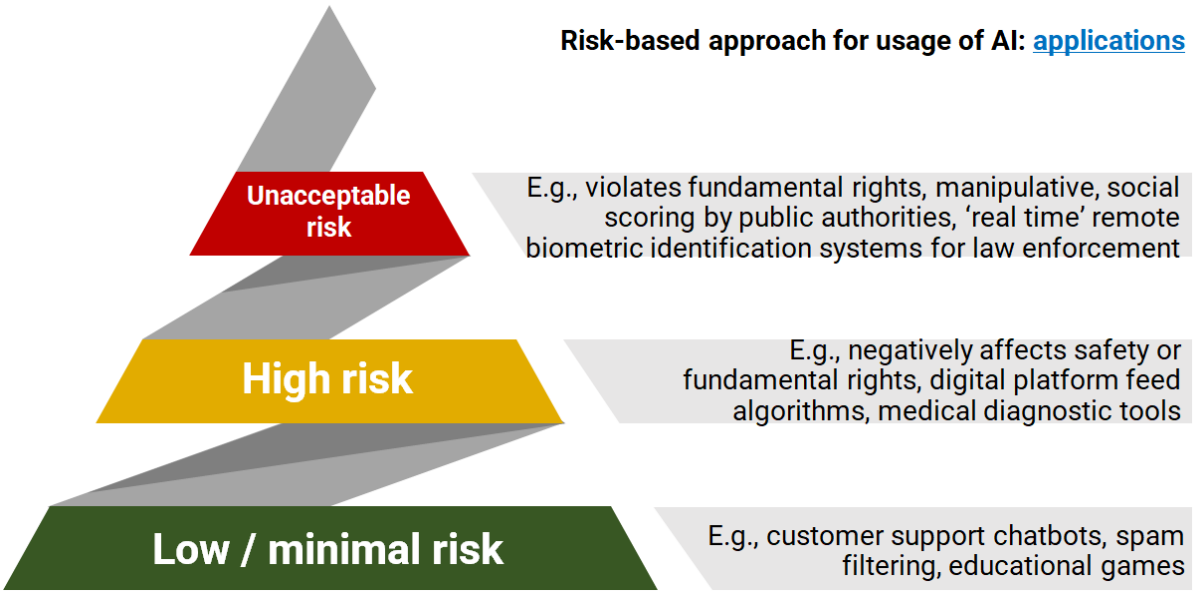

Figure 3: Our proposed risk-based framework for applications.

See our response to Question 10 for a comprehensive discussion of what applications should be considered "unacceptable risk".

See our response to Question 19 regarding reasons why LLMs and MFMs are in the "high risk" category.

See our response to Question 17 for comments on regulation of minimal-risk applications.

[59] Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (for simplicity, we refer to the proposed draft *EU AI Act* and not the most recent version adopted by the European Parliament on 14 June 2023 with amendments)

[60] Please note that most of such AI is only high-risk as an application, not as a technology. Therefore we do not propose to subject the development of social media feed algorithms to the same evaluations of dangerous capabilities as LLMs.

A risk-based approach for development of AI technologies

The EU classification is mostly concerned with applications or uses (e.g. social scoring, law enforcement, AI-enabled video games or spam filters). **We propose extending this framework to research and development of AI technologies** (Figure 4).

It is important to cover development because once a dangerous technology is developed, it can be hard to stop its proliferation and misuse. It is most practical to take a preventative approach towards the development of models with unacceptable risk. We also have an ethical responsibility not to allow a technology to grow beyond our understanding of it. We must design safety guardrails and cross checks at the earliest stages of development, such as detailed documentation and external reviews.

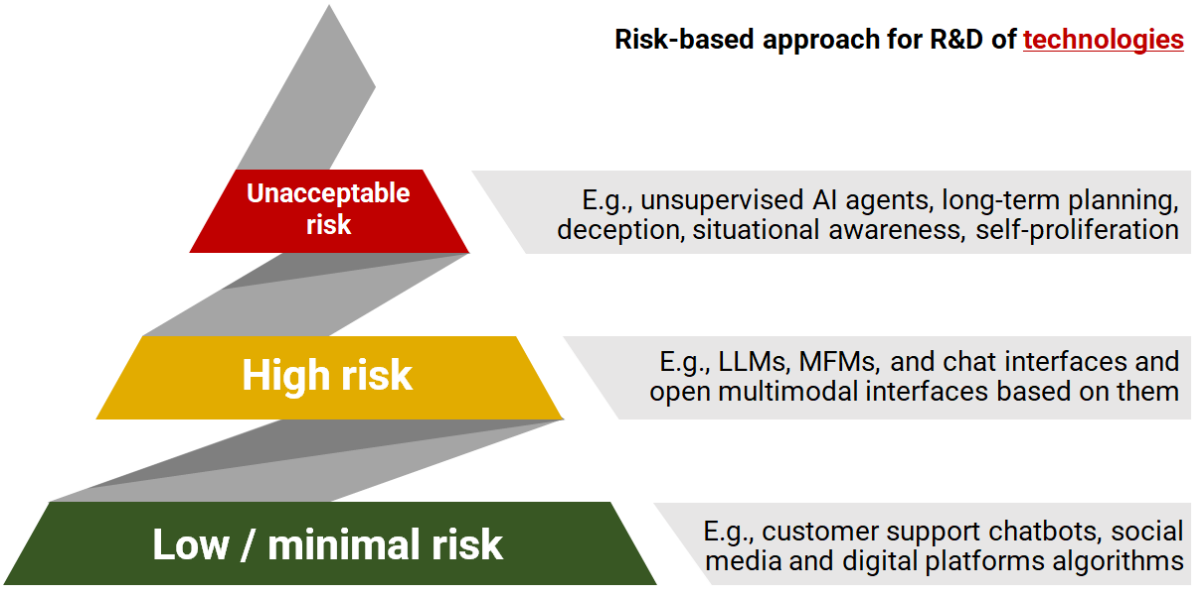

Figure 4: Our proposed risk-based framework for research, development, and deployment of AI (not to be confused with the risk-based framework for applications).

Unacceptable risk in development

We feel that the research and development of cutting edge, frontier AI models with the potential for dangerous capabilities should be prohibited in Australia. These models are harmful and pose a clear threat to people's safety, livelihoods and rights. As AI developers such as Anthropic make reference to AI models that will "automate large portions of the economy"[61], they pose a direct threat to people's livelihoods. This tier should also include performing training runs of models that exceed the computer required to run GPT-3. Google DeepMind provides an overview of dangerous capabilities[62] which we consider should fall in this tier.

See our response to Question 10 for a more comprehensive discussion of what technologies should be prohibited.

[61] "Anthropic's $5B, 4-year plan to take on OpenAI", Kyle Wiggers, Devin Coldewey, Manish Singh, TechCrunch (7 April 2023).
[62] "Model evaluation for extreme risks", Toby Shevlane, et al. (24 May 2023).

High risk in development

AI developers would need to apply for a licence for training and deployment of large general-purpose AI systems and comply with safety and reporting requirements (see Appendix A for details).

A model that has a chance of posing unacceptable risk should be considered high-risk. As such, the definition should cover several aspects:

- Compute used in training. We propose a cut-off of 10^{22} FLOP (or 10 times less than GPT-3)
- Breadth of capabilities and intended generality of applications
- Amount of training data (e.g. any model that is trained on more than 30 TB of text).

If any of three factors is present, the system should be considered "unacceptable risk".

The regulator enforcing this framework (in our proposal, an Algorithm Review Board) would maintain a database of AI systems and continuously review AI practices given the fast pace of AI development. Australia should also recognise licences issued in other countries with suitable regimes. The regulator would have surveillance powers to monitor AI data centres in Australia for unauthorised training runs and the power to issue large fines[63] and/or prosecute if these unauthorised runs occurred.

Evaluation for dangerous capabilities

A major risk with general-purpose AI systems is that its capabilities can be used for very harmful purposes that were not intended by the AI developer resulting in societal or third-party impact. This could encompass: social manipulation[64], skills in deception[65], use of AI in warfare[66], and cybersecurity threats[67]. AI developers, regulators and end users may not be aware of these dangerous capabilities, which could lead to overuse or reliance on AI which are harmful for society. This overuse or reliance could well come at the cost of beneficial AI.

The AI industry already conducts some safety evaluations of its technology, but the practice is not consistent, is rarely publicised and may not include testing for the very dangerous capabilities described above. Even where risks are discovered, the AI company would face huge financial pressure to bury the results and fix the problem quietly, which would increase the risk that a different company makes the same mistake. Existing laws may limit the misuse of AI for malicious intents such as cybercrime offences in the *Criminal Code Act 1995 (Cth)*, however there is no existing law that adequately addresses this risk.

---

[63] We propose percentage-of-worldwide-turnover fines e.g. 6% of global revenue for non-compliance with unacceptable risk requirements, 4% of global revenue for other regulatory requirements.
[64] Some AI products have the ability to systematically deceive human users. The CICERO program from Meta achieved tournament-level results at Diplomacy by lying to human players. LLMs display 'sycophancy', giving different answers to users depending on stated political background etc.
[65] "Update on ARC's recent eval efforts", The Alignment Research Center (17 March 2023): The Alignment Research Center found that GPT-4 could pretend to be a blind person to hire a human via an online job ad to pass the CAPTCHA test so that it could access the internet.
[66] "A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says", Joe Hernandez (1 June 2021): Autonomous drones have been deployed in Libya that may have already selected and killed combatants without human oversight; "An AI Just Beat a Human F-16 Pilot In a Dogfight — Again", Patrick Tucker, Science & Technology Editor, Defense One (20 August 2020): Ais have outperformed human pilots in F-16 simulations.
[67] "WormGPT: New AI Tool Allows Cybercriminals to Launch Sophisticated Cyber Attacks", The Hacker News (15 Jul 2023): A new generative AI cybercrime tool called WormGPT can launch sophisticated phishing and business email compromise attacks.

We propose that AI companies must conduct safety evaluations prior to deployment in Australia and that the results are disclosed to users. Google DeepMind's evaluation framework[68] can be used as a guide to develop design criteria to assess whether a model has certain dangerous capabilities (defined in the framework) and whether it has the ability to harmfully apply those capabilities. A formal reporting process to regulators for these evaluations will need to be established. These evaluations must also be externally audited.

By mandating evaluations, the objective is to incentivise AI developers to account for third-party impacts in the intense competition to develop more advanced AI, and foster demand for safer and better AI.

This recommendation would apply to high-risk, general purpose AI systems in Australia only. It is reasonable to expect AI developers to absorb this regulatory burden because they are well resourced to fund independent evaluations (being large tech corporations) and have the most information about this cutting-edge technology. They are also most likely to develop dangerous AI systems.

As AI technology is a global market, we realise this recommendation would be most effective at mitigating the negative externalities if it were adopted internationally. The Australian Government can play an important role in this by working with its trading partners and ensuring the upcoming AI safety summit in the UK is used to advocate for independent safety evaluations.

Low or minimal risk in development

Australian businesses are free to develop AI in this category without inhibitions.

[68] "Model evaluation for extreme risks", Toby Shevlane, et al. (24 May 2023).

Question 17. Elements of risk-based approach

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

Attachment C and the matrix presented in Box 4 in the discussion paper are appropriate for medium and high risk applications. However:

- Most proposed requirements for low-risk applications are impractical and unenforceable. Putting them into law may have a diluting effect on compliance with other requirements. We believe that disclosure of use of AI is the only appropriate obligation for low-risk applications. The other measures should remain as recommendations.
- They are relevant only to applications, not development. A different set of obligations is required for development (see Appendix A).

Question 19. Application of risk-based approach to general purpose AI systems

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

Using our proposed risk-based approach (outlined in Question 14) :

- Existing LLMs and MFMs are categorised as "high risk" technologies because they have high generality of capabilities and applications, require a lot of compute for training, and a lot of training data.
- Development of LLMs or MFMs using more than 10^{23} FLOP in compute should be considered as "unacceptable risk".

Question 20. Applicability and regulation of risk-based approach

"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?"

Risk-based approach must be regulated

The risk-based approach should be mandated through regulation, which applies to **both** public and private organisations, **both** developers (i.e. firms that research, develop, and provide access to AI models) and deployers (i.e. firms that are responsible for application of AI).

AI does not need any sandbox or testbed schemes. AI in various forms has been in use for many years in government and in business. The growth of the sector is not impeded by existing regulation. New regulation that we propose mostly affects high-risk applications, for which any relaxations of regulatory regimes would be inappropriate.

Voluntary schemes in relation to artificial intelligence are seen as failing both in the Australian context[69] and globally due to a variety of factors[70]. Furthermore, according to our exploratory poll[71], Australians overwhelmingly support government regulation of artificial intelligence.

Creation of an Australian Algorithm Review Board

There are different options for a new regulatory machinery that will enforce the risk-based approach in development and oversee some high-risk applications of AI. These include:

- Setting up a specialist regulatory division within ACMA. The ACMA's remit is communications and airwaves which makes it a natural fit for AI and it has previously explored the impact of AI on the communications and media markets[72]. ACMA has expertise in developing technical standards which may be helpful in developing capability in the technicalities of AI. As part of the Digital Platform Regulators Forum (alongside, the Office of the Australian Information Commissioner, and eSafety Commissioner and ACCC), ACMA is examining whether generative AI falls within its regulatory scope[73].

  A downside of placing the new functions with ACMA is that ACMA already has a significant scope of duties in other areas.

[69] "Australia's AI Ethical Framework: Another paper tiger?", Jodie Siganto, Privacy 108 (08 Jul 2021).
[70] "Principles alone cannot guarantee ethical AI", Brent Mittelstadt, Nature Machine Intelligence (04 November 2019): volume 1, pages 501–507.
[71] "Public opinion in Australia", Nik Samoylov, Campaign for AI Safety (12 July 2023).
[72] "Artificial intelligence in communications and media", The Australian Communications and Media Authority (July 2020).
[73] "Digital Platform Regulators Forum puts generative AI on agenda", The Australian Communications and Media Authority (4 July 2023).

- Setting up a specialist regulatory division within ACCC. Over the past few years, the ACCC has conducted three inquiries[74] into the practices of digital platforms. Its monitoring of digital markets and investigations has led to it developing institutional capabilities in regulating digital markets, which we feel can be extended to include AI technology. The ACCC has also conducted enforcement actions against big tech companies (the same companies that develop high risk AI systems) for anti-competitive behaviour.

  A downside of placing the new functions with ACCC is that regulation of AI requires machine learning technical expertise, which is not held within ACCC.

- Setting up a specialist regulatory division within the Department of Industry, Science, and Resources. The Department is a natural fit for AI regulation, but because it is also charged with delivering the Government's agenda for creating 1.2 million tech-related jobs by 2030, the AI ethics and safety regulatory agenda may come into conflict.

- Establishing a new regulator (the preferred option).

We support the establishment of a new regular, an Australian Algorithm Review Board, which should serve as a nexus regulator for high-risk development of AI as well as certain high-risk applications of AI, such as within select digital platforms.

We believe that the case for such a body existed before the advent of generative AI and was very well laid out in the News Corp Australia submission to the ACCC Digital Platforms Inquiry[75]. In the current context, an Australian Algorithm Review Board should be responsible for not only regulation of monopolistic digital platforms, but also for:

- Licensing of high-risk AI development,
- Enforcement of the ban on unacceptable risk development and applications,
- Oversight of high-risk AI applications in digital platforms.

Proposed functions of an Australian Algorithm Review Board

We propose the following specific functions of the new regulator:

- **Licensing of high-risk AI development:**
  - Detail the licensing regime and requirements for AI companies,
  - Establish a process for recognition of foreign-issued licences,
  - Manage the licensing process (whether the regulator themselves performs some of the licensing steps such as evaluations or requires that independent third parties do that can be decided at a later stage),
  - Issue penalties for non-compliance,

[74] "Digital Platforms inquiry 2017-19"; "Digital platform services inquiry 2020-25"; and "Digital Advertising Services inquiry 2020-21", Australian Competition and Consumer Commission.
[75] "Submission to the Australian Competition and Consumer Commission Digital Platforms Inquiry issues paper", News Corp Australia (2018).

    - Investigate companies suspected of non-compliance,
    - Conduct audits of compliance.
- **Enforcement of the ban on unacceptable risk development and applications:**
    - Monitor compute in Australia via:
        - Surveillance of power usage within data centres,
        - Establishing a know-your-customer (KYC) scheme for cloud compute providers,
        - Reporting of compute activities in data centres.
    - Track dealing in computer chips[76] that can be used in deep learning.
    - Fine and/or prosecute companies or individuals that attempt to develop AI with unacceptable risk.
- The functions for **oversight of high-risk AI applications in digital platforms** are as described in the News Corp Australia submission to the ACCC Digital Platforms Inquiry.

Proposed powers of an Australian Algorithm Review Board

From the above, we see the following powers of the new regulator:

- Administer new legislation on high-risk and unacceptable risk AI development
- Compel companies to submit information on AI development, compute, as well as manufacturing, import, export, and dealing in high-risk computer chips
- Investigate, issue fines and direct orders to companies
- Prosecute for non-compliance
- Refer matters to the Australian Federal Police
- Other powers described in the News Corp Australia submission to the ACCC Digital Platforms Inquiry.

Proposed structure and funding of an Australian Algorithm Review Board

We see the following divisions within the new regulator:

- AI development:
    - Regulatory policy
    - Licensing
    - Compliance and investigations

[76] See for reference "What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring", Yonadav Shavit (30 May 2023).

- Monitoring and oversight:
    - Chip monitoring
    - Compute monitoring
- Digital platforms
- Revenue-raising
- Corporate functions

One of the key benefits of putting together AI development and digital platforms is that it is possible to **levy a Compliance Fee** on dominant digital platforms and AI licensees that will fund the operations of the new regulator. There is a significant overlap between digital platforms and frontier AI developers already. The Compliance Fee can be levied from the inception of the regulator and should be sufficient to cover the expenses of the new body.

Staffing considerations

Recruiting staff with relevant expertise will be the biggest challenge in setting up the new regulator. However, because there are no major AI labs located onshore, there is no immediate urgency to fully staff all areas. Recruitment and set-up of the regulator can take several months. It is important that companies are not allowed to operate without a licence before the licensing apparatus is established.

Some of the areas of regulation required (such as monitoring of compute) are new, not only in Australia but globally. There are no established or credentialled experts in some of these areas. We recommend hiring a mix of people with expertise in regulation and risk management in high-risk industries (such as nuclear energy, defence, and biosecurity), machine learning, and AI ethics. The compliance and investigations division should have staff with experience in policing and cybersecurity.

Maintenance of the policy function of the Department

Notwithstanding the creation of an Australian Algorithm Review Board, the Department of Industry, Science, and Resources should maintain its policy function in relation to AI.

New power for the Minister for Industry and Science

The Minister for Industry and Science should have a new power to ban certain machine learning, computer programming, and chip-related (manufacture, supply, import, export) activities as and when they are identified as dangerous.

Role of IP Australia in model transparency

As per our response to Question 9, we believe that high-risk AI technologies must come with model cards that are accessible to users. IP Australia has established expertise in databases of bibliographic and technical information. IP Australia is already exploring issues at the intersection of IP and AI, in addition to examining the role of IP in the development

and adoption of AI systems. It is a natural fit to become a centre for disclosure of model cards, including housing the bibliographic references of high-risk models. AI developers should be able to fulfil their disclosure requirements via an online portal in an agile fashion.

The new function can be funded with new fees paid by applicants (AI companies). In line with the existing funding model of IP Australia, the new mandatory disclosures can help grow net income to the Australian Government.

Role of eSafety Commissioner

The enforcement of the new rules on disclosure of AI-generated content should be within the functions and powers of the eSafety Commissioner.

Appendix A: Safety and reporting obligations on AI developers in ‘high risk’ category

There are various existing proposals for safety requirements for AI developers. Given the magnitude of risk, we encourage the “Swiss Cheese” model[77] of risk management:

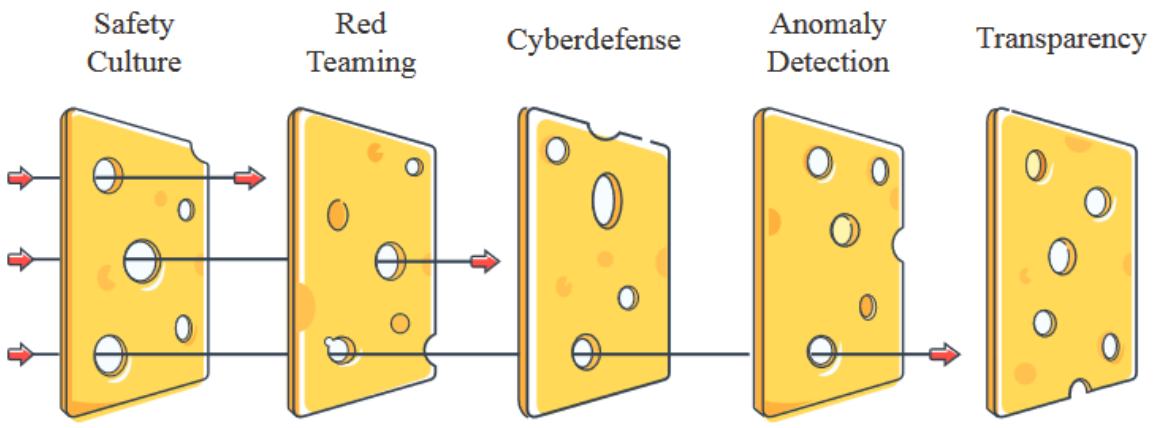

Figure 14 from “An Overview of Catastrophic AI Risks”: The “Swiss Cheese” model shows how technical factors can improve organisational safety. Multiple layers of defence compensate for each other’s individual weaknesses, leading to a low overall level of risk

Below is our proposal for an initial set of obligations on AI developers in ‘high risk’ category, which can be supplemented with additional requirements:

Pre-training authorisations

Pre-training model evaluations[78] will require AI companies to state the intended characteristics of AI models, their expected behaviours, and datasets used in training the data. These must be made publicly available 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 to unsafe model training and regulators need powers to prohibit the creation or fine-tuning of AI systems that are potentially dangerous.

Pre-deployment safety evaluations

The AI 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, these evaluations can be modelled based on evaluations run by the Alignment Research Center[79] and the framework proposed by Google[80]. These evaluations must 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.

[77] “An Overview of Catastrophic AI Risks”, Dan Hendrycks, Mantas Mazeika, Thomas Woodside (11 Jul 2023).
[78] “The Case for Pre-emptive Authorizations for AI Training”, Lennart Heim (10 June 2023).
[79] “Home” Page, ARC Evals.
[80] “Model evaluation for extreme risks”, Toby Shevlane, et al. (24 May 2023).

Safety committees

Licensees should be required to have safety committees that are similar to internal risk management committees at banks and other financial institutions. Safety committees must be responsible for AI safety of models and their implementations. We propose the following features:

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

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

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

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

e) at least a third of safety committee members are government appointed and remuneration is publicly funded and not tied to the licensee's financial performance.

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, the independence of the safety committees, etc);

b) the process of training and deployment of models;

c) the 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. To enable rapid implementation, an ecosystem of safety auditors and standard-setters should be mandated by the government and established by the 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[81] to businesses looking to purchase access to AI systems. This is documentation detailing a model's performance characteristics, training data (with comprehensive references), intended context of use and 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.

Mandatory technical AI safety research

a) Invest at least 50% of its aggregate research spend on advancement of safety, reliability, and explainability[82] 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 the deployment of unsafe systems.

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

Other measures

- Stringent cybersecurity protocols and compliance with ISO/IEC 27001 and SOC2 Type II.
- Restriction of end-user access to models via API (rather than provision of model copies

[81] "Model Cards for Model Reporting", Margaret Mitchell, et al. (14 Jan 2019).
[82] The other 50% can be spent on optimising foundation models up to GPT-3 capabilities for specific commercial or public applications.