Defining ‘high risk’ AI: Executive Summary

Good Ancestors is an Australian charity dedicated to reducing existential risk and improving the long-term future of humanity. We care about today’s Australians and we care about future generations. We believe that Australians and our leaders want to take meaningful action to combat the big challenges Australia and the world is facing.

The risks created by AI systems include societal-scale harms, and predicting which AI systems pose the greatest risks is challenging. Given the high stakes and uncertainty, AI regulation should recognise the different *scales* of AI risk, whilst being *flexible* in light of uncertainty.

The Canadian AIDA procedure for adding new ‘high risk’ AI use-cases is appropriately flexible. The Canadian approach allows the regulator to add new types of systems to the ‘high risk’ legislative category after considering certain factors, including the risk of adverse impacts and the extent of those impacts. The EU AI Act fails to incorporate sufficient flexibility. Given the uncertainty associated with AI capabilities and developments, this flexibility is desirable.

What matters to risk is the actual capabilities of an AI system, not the intent of its designers. Both the EU AI Act and the Canadian Act define ‘high risk’ AI through reference to the *intent* of the system provider. A definition based on system *capabilities* would be more appropriate.

The EU AI Act’s definition of ‘general purpose’ is better than the AIDA equivalent. The EU AI Act defines ‘general purpose systems’ via the *generality* of tasks the system can perform. This definition better targets the systems needing regulation than the Canadian definition, which focuses on the *number of fields* a system may be useful in.

The EU AI Act’s inclusion of a category for ‘general purpose systems that pose systemic risks’ best reflects the reality of AI risk. This category appropriately captures frontier AI development and allows regulation to target risks from new AI systems without overregulating systems we know are safe. The Canadian approach fails to do this as effectively.

Any organisation regulating AI should not also be tasked with fostering economic development. The Canadian Bill received substantial criticism for assigning regulatory tasks to the Department charged with supporting innovation and economic development. Safety and maximising productivity can be in actual and perceived tension. Australia should avoid conflicts between regulatory effectiveness and economic growth.

Defining ‘high risk’ AI:
Comparing Canadian and EU Approaches

The interim response to the ‘Safe and Responsible AI in Australia’ consultation highlighted the Government’s intent to adopt a risk-based legislative framework for regulating AI.[1] The interim response flagged the need for further work on ‘defin[ing] ‘high risk AI’ in an Australian context’.[2]

Successful regulation of high risk AI must align legislative categories with the risks that AI systems pose. This paper does not evaluate the types of obligations that should be placed on AI system providers. Rather, it focuses on how AI systems ought to be *classified* with reference to the Canadian and EU approaches. Good Ancestors intends to publish a future report on appropriate *obligations* for high risk and general-purpose systems in the Australian context.

This piece has three parts:

> **Part I** highlights technical research about the nature of the risks posed by AI systems. Specifically, risks vary in magnitude, and there is uncertainty about which systems will pose the greatest risks.
>
> **Part II** outlines the European Union’s (‘EU’) and Canada’s approach to defining and regulating ‘high risk’ AI.
>
> **Part III** highlights four core differences between the EU and Canadian approaches and assesses them against the principles outlined in Part I. Recognising that Australia can learn from the approaches taken by the EU and Canada, this piece proposes areas in which Australia can adopt different aspects of each jurisdiction's definitional regime.

[1] *Safe and Responsible AI in Australia Consultation: Australian Government’s Interim Response* (Report, Australian Government, 17 January 2024) 20 (‘*Interim Response*').
[2] Ibid.

Part I: The Nature of AI Risk

The capabilities of artificial intelligence systems have increased dramatically in the last decade.[3] Frontier AI systems can locate cybersecurity vulnerabilities,[4] solve complex mathematical problems,[5] and perform well on many other tasks.

At its most basic, risk is the product of **consequence** and **likelihood**.[6] To understand the potential risks posed by emerging technology, we need to consider the magnitude of an outcome's consequence and the likelihood of that outcome occurring.

A Variation in Magnitude

The potential risks of AI vary widely in potential magnitude.[7] Already, some harms from AI systems are ubiquitous. We are regularly distracted by algorithms designed to be addicting, and we may suffer privacy harms as information is harvested by content scraping algorithms. Other AI systems may lead to more serious harms. For example, the creation of deep fakes can increase the individual harm caused by cyber abuse.[8]

Beyond these forms of harm, a large portion of the expert AI community is concerned by 'algorithmic black swans'[9] – societal-scale risks created by AI that are difficult to predict.[10] Recent surveys of machine learning researchers show concern regarding 'bad' and 'extremely bad' outcomes from AI development.[11] Similarly, in 2023 the CEOs of the largest AI companies, as well as many leading AI researchers, stated that mitigating risks from AI should be treated with similar seriousness to pandemics and nuclear weapons.[12]

---

[3] Noam Kolt, 'Algorithmic Black Swans' (2023) 101(4) *Washington University Law Review* 1177, 1188.
[4] Richard Fang et al, 'Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities' (No arXiv:2406.01637, arXiv, 2 June 2024) <http://arxiv.org/abs/2406.01637>.
[5] Stanislas Polu et al, 'Formal Mathematics Statement Curriculum Learning' (No arXiv:2202.01344, arXiv, 2 February 2022) <http://arxiv.org/abs/2202.01344>.
[6] See, eg, David Farber, 'Uncertainty' (2011) 99 *Georgetown Law Journal* 901, 907–8; Mahler, Tobias Mahler, 'Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal' [2022] *Nordic Yearbook of Law and Informatics* 247, 257.
[7] This is acknowledged in the Australian Government's interim response and is part of the motivation for the stated intention to adopt a 'risk-based approach'; *Interim Response* (n 1) 4, 11.
[8] *Interim Response* (n 1) 11.
[9] Kolt (n 3) 1195.
[10] See, eg, the consistent concern regarding 'bad' or 'extremely bad' outcomes in surveys of AI researchers; Katja Grace et al, 'When Will AI Exceed Human Performance? Evidence from AI Experts' (No arXiv:1705.08807, arXiv, 3 May 2018) 4 <http://arxiv.org/abs/1705.08807>; Baobao Zhang et al, 'Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers' (No arXiv:2206.04132, arXiv, 8 June 2022) 2 <http://arxiv.org/abs/2206.04132>;
[11] Grace et al (n 10) 4; Zhang et al (n 10) 2; Michael et al (n 10) 2; Kolt (n 3) 18.
[12] 'Statement on AI Risk', *Centre for AI Safety* (Web Page) <https://www.safe.ai/work/statement-on-ai-risk>.

Although taxonomies of these societal risks are diverse, they can roughly be characterised into three categories.[13]

1. **Risks from the *misuse* of AI systems.** AI systems may be used by malicious actors to cause substantial harm. For example, research suggests that current or future AI systems may lower the knowledge barriers required to create biological weapons or launch cyberattacks, allowing relatively unsophisticated actors to cause substantial harm.[14]

2. **Risks from AI *accidents***. AIs deployed in high-stakes environments, such as infrastructure, cybersecurity, or military applications could malfunction, causing substantial harm.[15] This could include risks from 'misaligned' AI acting contrary to its operator's intent.[16]

3. ***Structural risks*.** AI systems could undermine societal institutions.[17] For example, AI may be used to create wide-spread disinformation, distorting individual values and destabilising organisations.[18] Research demonstrates that AI systems used by social media companies promote politically divisive content and influence voting trends.[19] Whilst this causes individual harms, it also risks substantial aggregate harm.[20]

Regulation needs to account for AI risks varying widely in magnitude. They range from individual harms (e.g. cyberbullying, privacy harms) to societal-scale harms (e.g. cyberattacks, biological weapons). Regulators have previously grappled with ubiquitous technologies that can cause appreciable harm, like cars or planes. Equally, regulators have grappled with constrained technologies that can cause catastrophic harm, like nuclear weapons or biotechnology. AI presents a **unique regulatory challenge**, being potentially ubiquitous whilst also being able to cause catastrophic harm.

---

[13] For examples of this taxonomy, see Toby Ord, Angus Mercer and Sophie Dannreuther, *Future Proof Report* (Centre for Long-Term Resilience, June 2021) 24 <https://www.longtermresilience.org/post/future-proof-report-2021>.

[14] William D'Alessandro, Harry Lloyd and Nathaniel Sharadin, 'Large Language Models and Biorisk' (2023) 23(10) *The American Journal of Bioethics* 115; Even more sceptical research emphasises their focus on "current systems". See, eg, Christopher Mouton, Caleb Lucas and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study* (Research Report, RAND Corporation, 27 January 2024) ('the potential for an unknown, grave biological threat propelled or even generated by LLMs cannot be ruled out").

[15] Kolt (n 3) 1191; Helen Toner and Zachary Arnold, *AI Accidents: An Emerging Threat* (Policy Brief, Centre for Security and Emerging Technology, July 2021) 7–10;

[16] Dan Hendrycks et al, 'Unsolved Problems in ML Safety' (No arXiv:2109.13916, arXiv, 16 June 2022) 8 <http://arxiv.org/abs/2109.13916>.

[17] Kolt (n 3) 1223; Toby Ord, Mercer and Dannreuther (n 13) 24.

[18] See, eg, Josh Goldstein et al, 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations' (No arXiv:2301.04246, arXiv, 10 January 2023) <http://arxiv.org/abs/2301.04246>; Kolt (n 3) 1226.

[19] Philipp Lorenz-Spreen et al, 'A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy' (2023) 7(1) *Nature Human Behaviour* 74; Robert Bond et al, 'A 61-Million-Person Experiment in Social Influence and Political Mobilization' (2012) 489(7415) *Nature* 295.

[20] Kolt (n 3) 1225; Jack Citrin and Laura Stoker, 'Political Trust in a Cynical Age' (2018) 21(1) *Annual Review of Political Science* 49, 50.

B Uncertainty and Unpredictability

Regulatory theorists recognise that many policy issues require making regulatory choices with limited information.[21] Research shows that uncertainty can lead to policymakers dramatically underestimating potential risks – especially the most extreme risks – of a particular regulatory challenge, such as those posed by climate change or pandemics.[22]

Similarly, the capabilities of AI models, and thus their risks, are uncertain.[23] Quantitative increases in the amount of data used to create an AI system can lead to unpredictable leaps in AI capabilities, a phenomenon called the '**unexpected capabilities**' problem.[24] Similar challenges arise from '**capability overhang**'.[25] Capability overhang is a phenomenon where existing AI systems are found to have capabilities that were previously thought to be beyond their range, but are discovered post-deployment through prompting or unique use.[26] For example, through chaining together instances of Chat-GPT4, researchers have found that it can exploit cybersecurity vulnerabilities, a previously unknown capability of current models.[27]

There is also uncertainty about how AI systems may operate in practice, a challenge called the '**deployment safety problem**'.[28] AI systems may seemingly be operating as intended, but once exposed to new scenarios,[29] act in unexpected ways.[30] Such issues have been demonstrated in 'low-risk' scenarios, such as game-playing AI's malfunctioning in unexpected ways.[31] However, the literature suggests this problem may only increase as AI systems are deployed in higher-stakes settings, such as public infrastructure.[32]

Collectively, these challenges make **anticipatory regulation** necessary because the onset of AI risks could be sudden and consequences potentially severe.

---

[21] See, generally, Farber (n 6); Cass Sunstein, 'The Limits of Quantification' (2014) 102(6) *California Law Review* 1369; Jonathan Masur and Eric Posner, 'Unquantified Benefits and the Problem of Regulation Under Uncertainty' (2016) 102 *Cornell Law Review* 87, 89.

[22] Farber (n 6) 907–19.

[23] Many scholars have noted this difficulty in AI regulation. For example, see, Margot Kaminski, 'Regulating the Risks of AI' (2022) 103 *Boston University Law Review* 1374, 1369;

[24] Markus Anderljung et al, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (No arXiv:2307.03718, arXiv, 7 November 2023) 9-10 <http://arxiv.org/abs/2307.03718>.

[25] Ibid 11.

[26] Ibid.

[27] Richard Fang et al, 'Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities' (No arXiv:2406.01637, arXiv, 2 June 2024) <http://arxiv.org/abs/2406.01637> 3.

[28] Anderljung et al (n 24) 10–13.

[29] Hendrycks et al (n 16) 8.

[30] Ibid.

[31] See, eg, 'Faulty Reward Functions in the Wild', *OpenAI* (Web Page, 21 December 2016) <https://openai.com/index/faulty-reward-functions/>; Joar Skalse et al, 'Defining and Characterizing Reward Hacking' (No arXiv:2209.13085, arXiv, 26 September 2022) <http://arxiv.org/abs/2209.13085>.

[32] Hendrycks et al (n 16) 8–10.

C Regulatory Design Principles for AI Risks

How should these characteristics of *uncertainty* and *variance in magnitude* inform decisions of regulatory design? Drawing on the relevant regulatory theory literature, Good Ancestors suggests two principles to guide such decisions:

1. ***Adaptability***. Regulatory schemes are often slow to adapt as technologies evolve, leading to both the *over-regulation* of beneficial technologies and the *under-regulation* of potentially harmful technologies.[33] Regulatory scholars argue that the regulatory process needs to be adaptive – able to change quickly and without the institutional inertia that can characterise legislative change.[34] This is particularly true for AI, where the rate of progress of AI systems is unpredictable and nonlinear. Thus, whilst certain use cases of AI may always be risky, **regulation must be *adaptable*, so it can impose or remove obligations as the capabilities and risks created by particular systems change over time**.

2. ***Proportionality***. Traditional forms of cost-benefit analyses may fail to appropriately account for uncertainty, leaving worst-case outcomes under-evaluated. Regulatory theorists have recognised that **regulation that fails to distinguish between the differences in risk may place overly burdensome obligations on systems that pose low to moderate levels of risk, or place overly *tolerant* regulations on systems that pose more substantial risks**.[35] Legislation must appropriately delineate between different risk profiles of AI, regulating both technologies that create individual harms, but also anticipate the potential for large-scale societal harm, and placing proportionate obligations respectively.

[33] Lori Bennear and Jonathan Wiener, *Adaptive Regulation: Instrument Choice for Policy Learning over Time* (Working Paper, Harvard Kennedy School, 12 January 2019) 1.
[34] For a small selection of the scholars that have argued as such with respect to different technologies, see generally, Lori Bennear and Cary Coglianese, 'Flexible Approaches to Environmental Regulation' in Michael Kraft and Sheldon Kamieniecki (eds), *The Oxford Handbook of U.S. Environmental Policy* (Oxford University Press) 582; Bennear and Wiener (n 33); Jody Freeman and Daniel Farber, 'Modular Environmental Regulation' (2005) 54(4) *Duke Law Journal* 796.
[35] Bennear and Wiener (n 33) 3-6.

Part II: International Approaches to AI Regulation

Both the EU and Canadian AI legislation have been highlighted as potential models for Australia,[36] both exist within similar political and judicial climates as Australia, and both their regulatory approaches are similar but distinct in informative ways. This section outlines each nation's (existing or proposed)[37] approach to AI regulation.

A The European Union's AI Act

The *EU AI Act* establishes a **risk-based regulatory framework** for AI systems throughout the European Union.[38] The Act places obligations on AI system 'providers' depending on a system's risk characterisation. The Act deems that some uses of AI systems pose *unacceptable risks*, such as the use of AI for 'social scoring' systems.[39] The most substantial legislative category is that of 'high risk' AI systems, defined in Article 6.

Article 6(1)(a) specifies that a system will be 'high risk' when it would be covered by EU product-safety harmonisation legislation, which covers products like machinery and medical devices.[40] Article 6(2) establishes that systems defined in Annex III are deemed 'high risk', which includes AI systems 'intended to be used' to select for educational and employment opportunities, to manage critical infrastructure, to assist public authorities in allocating benefits and to assist in the administration of justice, among a number of other uses.[41] Under Article 7, the European Commission – the executive arm of the EU – is empowered to modify Annex III. This can only be done where the AI systems are 'intended to be used' in one of the areas already outlined in Annex III, *and* where the system poses a risk equal to or equivalent to the systems referred to in Annex III.[42]

Beyond 'high risk' systems, the Act imposes separate, additional requirements for 'general-purpose AI models' ('*GPAI*'), and GPAI systems that pose 'systemic risks'.[43] The Act defines GPAIs as AI models that 'display significant generality' and are capable of 'competently performing a wide range of distinct tasks'.[44]

---

[36] *Interim Response* (n 1) 5
[37] At time of writing Canadian legislation is yet to pass Parliament.
[38] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 12/7 ('*EU AI Act*').
[39] Ibid art 5(1)(c), 5(1)(e).
[40] Ibid annex I.
[41] Ibid annex III art 1(a)-(c), 2, 3(a)-(c), 5(a)-(d).
[42] Ibid art 71(b).
[43] Ibid art 51(a)-(b), 55(1)-(3).
[44] Ibid art 3(63).

A GPAI is considered to pose 'systemic risks' if it has capabilities that 'match or exceed the capabilities recorded in the most advanced GPAI models',[45] a threshold that is presumed to be met if it was trained using a level of computing power set out in the legislation. This threshold may be changed by regulation.[46] As such, the providers of GPAI models with systemic risks face the most substantial burdens under the *EU AI Act*, being required, for example, to perform model evaluations and testing before deployment.[47]

B Canada's Artificial Intelligence and Data Act

The Canadian *Artificial Intelligence and Data Bill* (*AIDA*) tabled in 2022 as part of the broader *Digital Charter Implementation Bill 2022*,[48] proposes a similar risk-based framework for the use of AI in Canada.[49] Since the original draft legislation received criticism for delegating substantial definitional questions to regulations,[50] a number of amendments have been proposed by the Standing Commission of Industry and Technology (INDU), the text of which has since been published by the relevant Minister.[51] These suggested amendments represent the output of substantial engagement with stakeholders,[52] and present a considered risk-based framework that may inform Australian policy.[53]

In the *AIDA*, the major regulatory category is **'high impact'** AI systems, which is defined as any system of which 'at least one of the intended uses may reasonably be concluded to fall within a class of uses set out in [schedule 2]'.[54] Schedule 2 contains a similar list to the *EU AI Act*, including those intended for use in biometrics, for determinations of employment of recruitment, and within a court or administrative body.[55] However, the definition is *broader* than the *EU AI Act*, including use cases such as the 'moderation of content found on an online communications platform'.[56] Like the *EU AI Act*, these

[45] Ibid art 55(1)(a), 3(64).
[46] Currently, the threshold is set at 10^25 FLOPS ('floating point operations per second'). Ibid art 51(2), 51(3).
[47] Ibid art 55(1)-(3).
[48] Digital Charter Implementation Bill, C 2022, C-27, pt 3.
[49] 'The Artificial Intelligence and Data Act (AIDA) – Companion Document', *Canadian Government* (Web Page,13 March 2023) <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document> ('*AIDA Companion Document*').
[50] See, eg, Teresa Scassa, 'Regulating AI in Canada: A Critical Look at the Proposed Artificial Intelligence and Data Act' (2023) 101(1) *The Canadian Bar Review* 1, 7.
[51] François-Philippe Champagne, *Proposed Amendments to the Artificial Intelligence and Data Act* (Report to the Standing Committee on Industry and Technology, 28 November 2023) 1 ('*Proposed Amendments to the AIDA*') <https://www.ourcommons.ca/content/Committee/441/INDU/WebDoc/WD12751351/12751351/MinisterOfInnovationScienceAndIndustry-2023-11-28-Combined-e.pdf> When referring to *information* on the process provided by this letter, this piece refers to page numbers. When referring to the compilation of proposed amendments, this piece will refer to the section number given in the document. When the relevant sections are not included in these proposed amendments but remain unchanged, this piece will refer to the Digital Charter Implementation Bill, C 2022.
[52] *Proposed Amendments to the AIDA* (n 51) 2.
[53] *Interim Response* (n 7) 5, 14, 24.
[54] *Proposed Amendments to the AIDA* (n 51) s 38.
[55] Ibid sch 2.
[56] Ibid.

categories can be changed. However, the process under *AIDA* is less restrictive than the *EU AI Act*, only requiring the Governor-in-Council (GIC) to 'take into account' four prescribed factors, including the risk of adverse impacts and the extent of those impacts.[57] In practice, the GIC is likely to follow the advice of the AI Data Commissioner, a position created by *AIDA*, who would report to the relevant Minister.[58]

The *AIDA*, similarly to the *EU AI Act*, places additional obligations on the providers of GPAIs, which it defines as any system 'designed for use […] in many fields and for many purposes', including those not contemplated during development'.[59] The providers of GPAI's must follow a variety of requirements largely to be set out in the Bill by regulations.[60] However, unlike the *EU AI Act*, the AIDA does not contain a further legislative category for GPAIs with 'systemic risks'.[61]

[57] Ibid ss 36.1(2).
[58] Ibid ss 33(1)
[59] Ibid ss 39. For clarity, the relevant definition is found at page 19 of the document contained the proposed appeals.
[60] Ibid ss 7(1)(a)-(i).
[61] *Interoperability Comparison between the Proposed Artificial Intelligence and Data Act and the European Union's Draft Artificial Intelligence Act* (Report, Osler Hoskin & Harcourt LLP, 7 February 2024) 7 <https://www.accessprivacy.com/AccessPrivacy/media/AccessPrivacy/Content/Comparison-of-AIDA-and-EU-AI-Act-key-provisions.pdf>.

Part III: Comparing Canadian and EU Approaches

Canada and the EU take similar approaches to defining AI categories. The AIDA specifically seeks to enable 'interoperability' with the *EU AI Act*.[62] However, each differs in important ways. We assess four key points and recommend which approach Australia should adopt.

A Limitations on Adding 'High Risk' AI Use Cases

Article 7(1)(a) of the *EU AI Act* requires that any additions to the list of 'high risk' systems in the Act must be 'intended to be used' in one of the areas already covered by Annex III.[63] That is, the Act assumes the list of 'high risk' categories provided is already comprehensive and complete, with the Commission's ability being limited to the creation of new *subcategories* of high risk AI.[64] The Canadian approach allows the GIC to add new high risk use cases *without* requiring alignment with the established high risk uses already specified in legislation.[65] As such, the Canadian AIDA procedure for adding new 'high risk' use-cases provides the desirable flexibility discussed Part I(C).

The EU's narrow approach may partly be explained by the EU's broader legislative context, where the 'new legislative framework—the EU's approach to harmonising conditions for product safety—which explicitly provides an *alternative* avenue for a system to be classified as 'high risk' under the Act,[66] and potentially minimising the need for Annex III to be broadened later. However, in a legislative context that lacks such a broader harmonising framework, like Canada and Australia, this presents a restrictive limitation that unnecessarily limits the creation of new 'high risk' classifications without legislative change. **Given fast and unpredictable changes in the AI landscape and the importance of regulatory adaptability, the slow pace of legislative change, article 7(1)(a) of the *EU AI Act* is undesirable.**

[62] *AIDA Companion Document* (n 49) 1.
[63] *EU AI Act* (n 38) art 7(1)(a).
[64] *People, Risk and the Unique Requirements of AI: 18 Recommendations to Strengthen the EU AI Act* (Policy Brief, Ada LoveLace Institution, 31 March 2022) 15 <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Policy-briefing-People-risk-and-the-unique-requirements-of-AI-18-recommendations-to-strengthen-the-EU-AI-Act.pdf>.
[65] *Proposed Amendments to the AIDA* (n 51) ss 36.1(1)-(2).
[66] *EU AI Act* (n 38) art 6(1)(a)-(b).

One objection to the Canadian approach is that the decision-makers under the Act—the AI Commissioner and the Minister for Industry and Science–[67] who determine when to introduce new 'high risk AI' classification are both located within the department charged with 'supporting innovation and economic development'.[68] The goals of safety and maximising productivity can be in tension. As the *Organisation for Economic Co-operation and Development* Report on the *Best Practice Principles for Regulatory Policy* notes, **the assignment of a regulator to both industry and regulatory functions can not only 'reduce the regulator's effectiveness in one or both functions', but also 'fail to engender public confidence' in the relevant regulator**.[69] On this basis, experts have suggested that the AIDA 'depart[s] from well-established principles of regulatory independence'.[70]

These critiques go to the *institutional implementation* of Canada's definition of high risk AI, rather than the flexibility it provides. Such concerns could be addressed with appropriate institutions without resorting to the comparative inflexibility of the EU's approach. For example, in Australia, the *Therapeutic Goods Administration* (*'TGA'*) – empowered by the *Therapeutic Goods Act 1989* (Cth) – is tasked with regulating and categorising medical devices in Australia.[71] Unlike the *AI Commissioner* under the *AIDA,* the TGA is not also tasked with improving the economic efficiency of the Australian healthcare system.[72] Operating within a flexible legislative framework that places substantial reliance on regulations, the TGA is generally considered successful.[73] In light of this, **Australia should adopt the Canadian approach by allowing modifications to the categories of 'high risk' without an equivalent restriction to Article 7(1)(a) of the *EU AI Act* but mitigate the shortcoming by ensuring the regulator is separate from the Department of Industry**. This approach better reflects the principle of adaptability discussed in Part I, allowing legislation to *adapt* to developments and changes in AI capabilities.

---

[67] Digital Charter Implementation Bill, C 2022, C-27, s 33.

[68] For a description of the role of the relevant Department, see 'Innovation, Science and Economic Development Canada' *Government of Canada* (Web Page, 17 July 2024) <https://ised-isde.canada.ca/site/ised/en>; Scassa (n 50) 12.

[69] *The Governance of Regulators* (Report, Organisation for Economic Co-operation and Development, 29 July 2014) 34 <https://www.oecd-ilibrary.org/governance/the-governance-of-regulators_9789264209015-en>.

[70] Andrew Clement, 'AIDA's "Consultation Theatre" Highlights Flaws in a So-Called Agile Approach to AI Governance', *Centre for International Governance Innovation* https://www.cigionline.org/articles/aidas-consultation-theatre-highlights-flaws-in-a-so-called-agile-approach-to-ai-governance/.

[71] For a description of the role of the *TGA*, see further, Rosalind Hewett, Rebecca Storen and Emma Vines, *Therapeutic Goods: A Quick Guide* (Research Report, Parliamentary Library, 3 May 2022) 1.

[72] Ibid 1-3.

[73] See, eg, *Therapeutic Goods Administration Performance Report 2022-23* (Performance Report, Department of Health and Aged Care, July 2023) 11; See also, Peter Bragge, 'Think the Therapeutic Goods Administration is too conservative? Think again', *Monash Health and Medicine* (Web Page, 4 February 2022) <https://lens.monash.edu/@medicine-health/2022/02/04/1384422/think-the-therapeutic-goods-administration-is-too-conservative-think-again>

B Role of 'Intent' in Defining 'High Risk' AI

Both the *AIDA* and the *EU AI Act* refer to 'intent' in their definitions of 'high risk' or 'high impact' AI systems. Under the *AIDA*, a system will be 'high impact' where "at least one of the **intended** uses may reasonably be concluded" to fall within a specified list.[74] Similarly, under the *EU AI Act*, a system will be 'high risk' where it is **'intended** to be used' in any of the specified Annex III categories.[75] Under the *EU AI Act,* once a system is classified as 'high risk' the model creator is required to take actions to prevent 'reasonably foreseeable' misuse.[76] However, these obligations are only imposed once a system is already classified as 'high risk'.[77] That is, they only apply when it has already been determined that the system was *intended to be used* in a designated 'high risk' use case.[78]

The problem with this regulatory focus on *intent* is that it does not impose obligations on models that may not be *intended* to be used in a 'high risk' way, even where they can be used (perhaps with simple alterations or 'jailbreaks') by third-parties in dangerous ways. For example, a narrow system that was created with the intention of inventing useful medicines could be easily altered by a third party to create novel pathogens instead.[79] Importantly, such a system *may not* be defined as 'high risk', because the creators did not *intend* for it to be used in a 'high risk' way, despite the system's actual capabilities posing risks to public safety. As such under the EU AI Act and AIDA, because the system would not be classified as 'high risk', its creators would likely not be under obligations to take actions to prevent such misuse.

Of course, consideration of 'intent' is useful. If a developer intends their system to be used in a high risk way, this should be *sufficient* to find that their system falls within the 'high risk' regulatory category. However, intent cannot stand alone as the only factor. A system should also be definable as 'high risk' if it is *reasonably foreseeable* that the system could be used in a 'high risk' way.

The precise meaning of 'reasonably foreseeable' could be given clarity and precision through reference to the results of benchmarking capability elicitation, third party verification services, and the provision of assessment criteria set out in regulations to be updated as required. Such an approach is already used in the *EU AI Act* for determining the precise obligations that are placed on 'high risk' AI system providers,[80] and there is a

---

[74] *Proposed Amendments to the AIDA* (n 51) s 38.
[75] *EU AI Act* (n 38) annex III art 1(a)-(c), 2, 3(a)-(c), 5(a)-(d).
[76] *EU AI Act* (n 38) art 9(2)(a).
[77] Ibid art 9(1)).
[78] Ibid.
[79] See the discussion of 'dual use' medical drug discovery use cases in Fabio Urbina et al, 'Dual use of Artificial-Intelligence Powered Drug Discovery' (2022) 4 *Nature Machine Intelligence* 189.
[80] See generally, ibid art 40-49.

growing focus on such services and criteria by international 'AI Safety' bodies.[81] Ultimately, **we recommend that any Australian AI safety legislation define 'high risk AI' through a *two-limbed* approach. A system ought to be classified as a 'high risk' system if the system provider intends to use it in particular designated 'high risk' use-cases - following the *EU AI Act* and the *AIDA,* but also if it is *reasonably foreseeable* the system be used in a 'high risk' way.** Regulation should allow for the listing of specific capability benchmarks in regulation to identify capabilities that would make a model 'high risk'.

C Definition of 'General Purpose' AI Models

Recognising the *unexpected capabilities* and *deployment safety* problems associated with generalist AI systems discussed in Part I**,** both the *EU AI Act* and the *AIDA* impose additional obligations on the providers of 'general-purpose systems'. The *EU AI Act's* definition focuses on the model's 'performance in [multiple] distinct **tasks**' – which can be measured through benchmarks and technical procedures that may become a part of the standard-setting associated with the Act.[82] *AIDA's* definition focuses on the different '**fields**, purposes and activities' in which an AI may be used, including those 'not contemplated during the system's development'.[83]

The *AIDA's* definition is substantially broader than the EU AI Act, because it includes systems that can only complete one distinct task*,* if that *task* is useful to a variety of fields. For example, an AI system created to forecast weather may *technically be a narrow* system, but it could be used in a variety of fields and activities (scheduling, agriculture, etc).[84] Under the *EU AI Act*, this would not constitute a GPAI, as it could not perform well in a 'number of distinct tasks'. However, under the *AIDA* it would likely be regulated as a general-purpose model, as it is useful for several different 'fields', despite it only having a limited range of capabilities. Similar arguments could be made for other, task-specific AI models where the relevant task is applicable over many fields or activities, such as AIs created to check grammar,[85] or AIs used to create music.[86]

The issue with AIDA's approach is that it lacks the specificity required for placing proportionate obligations on AI providers, undermining the broader scheme of legislative characterisation. Vital aspects of the obligations on GPAI's are left to future regulations.[87]

---

[81] See, eg, National Institute for Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (Report, June 2024); UK AI Safety Institute, 'Advanced AI Evaluations at AISI, webpage (20 May 2024) <advanced-ai-evaluations-may-update>.'

[82] *EU AI Act* (n 38) art 3(64);

[83] *Proposed Amendments to the AIDA* (n 51) s 39.

[84] *Interoperability Comparison between the Proposed Artificial Intelligence and Data Act and the European Union's Draft Artificial Intelligence Act* (n 61) 5.

[85] For just one example of an 'AI based' grammar checker, see, eg, 'Grammarly: Free AI Writing Assistance' (Web Page) <https://www.grammarly.com/>.

[86] 'SoundDraw', an AI music creation company, specifically advertises the fact that its service could be used for a variety of different purposes; 'Soundraw' (Web page) <https://soundraw.io>.

[87] *Proposed Amendments to the AIDA* (n 51) ss 36.1(1)-(2).

These AIDA regulators are tasked with making regulations for a category that is potentially too wide – covering models that range from the *frontier*, cutting edge of AI development, to grammar checkers and other similarly narrow systems. Regulators will face the challenge of either placing overly onerous obligations on narrow models or placing overly lax obligations on general models. That is, the *AIDA* definition of GPAIs lacks the level of *fine-grained nuance* that is needed for the placing of proportionate obligations on model providers. **We recommend that future Australian legislation follow the *EU AI Act's* approach in defining GPAI with respect to performance in distinct tasks, rather than applicability in different fields.** This should be supported by reference to capability benchmarks rather than designer intent.

D Inclusion of 'Systemic Risks' as a Legislative Category

All general-purpose AI models do not pose the same level of risk. The more advanced a general-purpose model, the more likely it is to have capabilities that make it dangerous. Further, the *'unexpected capabilities'* and *'capabilities overhang'* problems are most apparent for the most advanced AI systems, which are most likely to have unexpected capabilities.

This distinction can be illustrated through a comparison to GPT-3 and more recent models, like GPT-4.[88] Both would constitute a GPAI for the purposes of the *AIDA* and the *EU AI Act.* However, they present substantially different risks. Previous models like ChatGPT-3 are trained on less data and computational power, resulting in less powerful capabilities, making the system less dangerous in the hands of bad actors.[89] Further, GPT-3 has been available to the public for longer, reducing the possibility of 'capabilities overhang'.[90] On the other hand, risks are exacerbated for models on the 'cutting edge' of AI technology, where the systems are likely to be *more* powerful, and where the relevant capabilities have not yet been reliably evaluated.[91]

However, only the *EU AI Act* recognises this delineation in risk, with its inclusion of a separate legislative category in the *EU AI Act* for GPAI models that pose 'systemic risks'. Importantly, the *EU AI Act's* definition of 'systemic risks'[92] is adaptable and can be

[88] For a technical comparison of the two models, and to see the advancements that ChatGPT4 made compared to the previous model compare the initial two technical reports released for each model; Tom Brown et al, 'Language Models Are Few-Shot Learners' (No arXiv:2005.14165, arXiv, 22 July 2020) <http://arxiv.org/abs/2005.14165>.
[89] Anderljung et al (n 24) 9.
[90] Ibid.
[91] Ibid.
[92] There are two ways for a general purpose system to be defined as posing 'systemic risks' in the *EU AI Act*. **First**, if it has "high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks" (art 51(1)(a)). This definition will be presumed to be fulfilled when the system is trained on a certain amount of *computational power* (art 51(2)). This is initially set at 10(^25) FLOPS, but this amount can be changed "in light of evolving technological developments" (art 51(3)). **Second,** a general purpose system may pose 'systemic risks' if the Commission decides it has capabilities *equivalent* to those described in art 51(1)(a) (art 51(1)(b)). This is determined by considering a

changed to reflect the movement of the 'cutting edge', acknowledging that AI systems will improve both with improvements in computational power and algorithmic efficiency.[93] However, under the *AIDA,* the Act only provides a single legislative category – 'general purpose models', under which all providers will face the same obligations. The EU's approach in having a further legislative risk category allows for more fine-grained delineations in the obligations placed upon frontier and non-frontier model providers, better reflecting the principle of proportionality discussed above. **We recommend that Australia follow the EU approach and include 'systemic risks' as a separate legislative category of 'general purpose AI' and adopt a flexible approach to defining 'systemic risks'.**

Conclusion

Australian AI regulation should be sufficiently proportionate to recognise the different scales of AI risks, whilst being sufficiently adaptable to reflect the uncertainty associated with AI capabilities. This piece assesses three differences between the EU AI Act and the AIDA, ultimately providing recommendations on how Australia should adopt different aspects of these nations' legislative regime. The key recommendations are below.

number of criteria, including the size of the model, the quality and size of its training data, the capabilities of the models, the number of end users, and the input/output modalities of the system (Annex XIII).

[93] *EU AI Act* (n 38) art 51(1)-(3).

Summary of Recommendations

Recommendation 1: *Ensure flexibility in the definition of 'high risk' AI* [**III(A)**].	Australian AI legislation should not adopt an equivalent to Article 7(1)(a) of the *EU AI Act.* Australia should follow the *AIDA* in allowing the relevant regulator to add new 'high risk' use cases beyond those initially set out in legislation.
Recommendation 2: *Avoid tasking the relevant AI regulators with goals that conflict with safety* [**III(A)**].	Australian AI regulation should avoid the approach of the AIDA and *not* assign the function of defining which systems are 'high risk' or 'general' to bodies also tasked with maximising economic output or productivity. A separate body or position tasked solely with ensuring the *safe* and *beneficial* development of AI should be established to monitor the application of Australian AI regulation and its' risk categories.
Recommendation 3: *Create a definition of 'high risk' AI that considers both the intent of the model provider and the possibility of dangerous capabilities that the model provider did not intend* [**III(B)**].	Australian AI regulation should not follow the approach of *EU AI Act* and the *AIDA* in *requiring* that, for a system to be 'high risk', its providers 'intended' to use it in a high risk use case. Such an intent should be *sufficient but not necessary* to classify a system as high risk. An alternate definition based on the *reasonable foreseeability* of a system being used in a high risk way should be provided.
Recommendation 3: *Adopt a precise definition of 'GPAI'* [**III(C)**].	Australian AI legislation should define 'general-purpose models' according to their *capabilities in tasks* – following the *EU AI Act* - rather than their *usefulness in different fields*, as found in the AIDA.
Recommendation 4: *Adopt a legislative category that recognises the greater risks posed by frontier' GPAI models than non-frontier models* [**III(D)**].	Australia should adopt an equivalent to articles 51, 55 and 56 of the *EU AI Act,* which lay out the relevant additional obligations and classification requirements for GPAI's that pose 'systemic risks.'

Good Ancestors thanks the community groups and volunteers who support its work on AI. Particular thanks go to Mr Daniel Marns for his substantial contribution to this paper.