Good Ancestors Policy

Good Ancestors is an Australian charity dedicated to improving the long-term future of humanity. We care about today's Australians and future generations. We believe that Australians and our leaders want to take meaningful action to combat the big challenges Australia and the world are facing. We want to help by making forward-looking policy recommendations that are rigorous, evidence-based, practical, and impactful.

Good Ancestors has been engaged in the AI policy conversation since our creation, working with experts in Australia and around the world while connecting directly with the Australian community.

Good Ancestors is proud to help coordinate Australians for AI Safety.

Our thanks go to the volunteers who provided input to this submission and who care so passionately about being good ancestors to future generations of Australians.

Recommendations

1. The specifics of generative AI behaviours, capabilities and supply chains mean that a technologically neutral approach may not provide practical consumer protections to Australians. To achieve the overall goal of reducing unsafe products in the market, policymakers must engage with the unique regulatory challenges that emerge from the specific nature of AI technology.

2. Policymakers should not focus on the adequacy of regulations for today's AI systems, but forecast forward to AI systems we should expect in coming years.

3. Australian AI deployers and users will often have limited ability to understand and control the risks of AI systems. Effective regulation must ensure that AI model and system developers are incentivised to take safety engineering seriously. This includes developing the tools and capabilities necessary to make models safe and to empower others in the supply chain to do the same.

4. The potential consequence of an AI accident greatly exceeds the potential consequences from typical consumer goods. The application of consumer law to AI must account for this difference in risk.

5. Australia should establish an AI Safety Institute with functions relating to understanding technical AI safety, communicating technical information to government, and facilitating cooperation between government, academia and industry.

6. Given practical pressure on businesses to adopt AI even when they cannot assess and mitigate risks, an effective regulatory regime must be able to identify risky behaviour in most cases and impose sufficient consequences to discourage it practically.

7. Australian consumer law should contribute to global AI safety norm-building by ensuring that those best able to mitigate risks are incentivised to take meaningful action.

8. Australian consumer law should contribute to global AI safety norm-building by encouraging AI developers to invest in safety. AI models and systems should meet safety expectations consistent with other advanced technology, like aviation.

9. Model Cards and Responsible Scaling Policies provide a roadmap that could be adapted for the context of consumer law.

Table of contents

Previous submissions

In the interests of brevity, Good Ancestors recommends that the Treasury reads previous relevant submissions, including:

- Good Ancestors' submissions to the two consultations in the Safe and Responsible AI series,
- Good Ancestors' submission to the Senate's Inquiry on Adopting AI,
- Good Ancestors' submission to NSW Parliament's Inquiry into AI, and[1]
- Shine Lawyers' submission, developed with Campaign for AI Safety, to the initial Safe and Responsible AI consultation.[2]

[1] **Good Ancestors Policy**, *AI Safety Overview* https://www.goodancestors.org.au/ai-safety.

[2] **Shine Lawyers**, *Submission to the Department of Industry, Science and Resources: Campaign for AI Safety* (Atanaan Ilango and Dr Benjamin Koh, 26 July 2023) https://www.shine.com.au/media/submissions/campaign-for-ai-safety.

Australian Consumer Law, emphasising a technology-neutral approach, is insufficient to protect Australians from AI risks

The overall objective of Australian consumer law is to enhance the welfare of Australians through the promotion of competition and fair trading and provision for consumer protection.[3] Key to the strategic agenda is the reduction of unsafe products and services in the Australian market.[4]

Today, Artificial Intelligence (AI) is on track to dramatically increase the number of unsafe products and services in the Australian market. Because of the nature of AI, including "black box" capabilities and the global supply chain connecting model and system developers and deployers that ultimately result in a particular service being on the market in Australia, a "technologically neutral" approach is unlikely to provide practical protections to Australians.

Good Ancestors' submission to the recent Department of Industry, Science and Resources "Mandatory Guardrails" consultation argued that AI presents a unique regulatory challenge that requires novel approaches:[5]

> *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 or medical devices. 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.*

Recommendation One: The specifics of generative AI behaviours, capabilities and supply chains mean that a technologically neutral approach may not provide practical consumer protections to Australians. To achieve the overall goal of reducing unsafe products in the market, policymakers must engage with the unique regulatory challenges that emerge from the specific nature of AI technology.

[3] *Competition and Consumer Act 2010* (Cth) s 2.

[4] **Legislative and Governance Forum on Consumer Affairs** with **Consumer Affairs Australia and New Zealand**, *Strategic Agenda 2018-2022: An Integrated and Harmonised Approach to Consumer Choice and Protection* (August 2018). https://consumer.gov.au/sites/consumer/files/2018/11/FINAL-Strategic-Agenda-2018-2022.pdf.

[5] **Good Ancestors Policy**, *AI Safety Overview* https://www.goodancestors.org.au/ai-safety.

The pace of change in AI capabilities and risks demands novel and ambitious thinking

GAP's previous submissions, in particular to the Australian Senate's Inquiry on Adopting AI, set out detailed evidence about the rapid growth in AI capability and reasons to believe that financial investment, larger data sets, increased compute, and algorithmic efficiency gains will drive significant growth in AI capability (i.e., the complexity of tasks that AI can perform, and the attendant risks), at least until the end of the decade.[6]

Recommendation Two: Policymakers should not focus on the adequacy of regulations for today's AI systems, but forecast forward to AI systems we should expect in coming years.

'Black box' AI systems challenge the ability of businesses and consumers to make informed choices about AI risk exposure

The discussion paper is right to reflect on the "black box" nature of AI systems – often referred to as "opacity" in the technical literature. A core challenge is that even AI developers don't know how advanced models work. This opacity is in contrast to other software which is designed for a specific function. We cannot generalise a regulatory stance from existing software to generative AI.

Sam Altman, CEO of OpenAI, has conceded that "interpretability" or "explainability" (technical methods to resolve opacity) remain unsolved problems and that OpenAI considers that it does not need to solve them to keep releasing more advanced AI models.[7]

Relevant to consumer law - opacity limits the ability of deployers (i.e. Australian businesses), users and regulators to understand and address AI risks. Indeed, the billion-dollar businesses that recruit the world's best talent can't reliably anticipate AI behaviours and capabilities[8] or effectively manage resulting risks. Policymakers cannot expect that non-experts across diverse fields will be able to understand and manage risks from AI.

[6] **Epoch AI**, *Key Trends and Figures in Machine Learning* (Online, 2023) https://epochai.org/trends.
[7] Rachel Curry, 'Sam Altman Says OpenAI Doesn't Fully Understand How GPT Works Despite Rapid Progress' (Online, 30 May 2024) *Observer* https://observer.com/2024/05/sam-altman-openai-gpt-ai-for-good-conference/.
[8] Stephen Ornes, 'The Unpredictable Abilities Emerging From Large AI Models' (Online, 16 March 2023) *Quanta Magazine* https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/.

The attributes of AI represent a significant shift from other products. General consumer products have risks that we can reasonably expect consumers of those products to understand and grapple with. Toys have choking hazards that carers can reasonably understand and mitigate when supported with adequate product information. Specialised products with hard-to-intuit risks are typically intended for expert or professional users.

Advanced AI is on track to undermine that distinction. AI is on track to be a powerful tool with risks and mitigations that experts struggle to understand and implement at the same time as being intended for use by almost everyone.[9] If AI was "sold in a box" with information about its hazards, there could never be enough space for the label. MIT is operating an AI Risk Repository which, at the time of writing, categorises over 700 AI risks.[10] Many of these risks are highly unintuitive – ranging from AI generating biased or toxic content due to poor quality training data, to trying to persuade or mislead users, to failures of capability and robustness when making decisions in novel environments.

Experts argue that these unique features point to the need for an approach where **Government regulation forces AI companies to internalise risks**. Slattery et al. say:[11]

> *AI's decision-making… is often unpredictable, opaque, and involves complex interactions between millions of parameters. This complexity makes understanding how an AI arrived at a decision, and consequently who is responsible for the consequences of that decision, very difficult.* ***In the absence of a regulatory or legal incentive to take safety engineering seriously, developers may release poorly designed AI systems, and people harmed by those systems may be left without recourse.***

Recommendation Three: Australian AI deployers and users will often have limited ability to understand and control the risks of AI systems. Effective regulation must ensure that AI model and system developers are incentivised to take safety engineering seriously. This includes developing the tools and capabilities necessary to make models safe and to empower others in the supply chain to do the same.[12]

[9] Connor Leahy, Gabriel Alfour, Chris Scammell, Andrea Miotti and Adam Shimi, *The Compendium* (Version 1.0.2, 11 November 2024) https://www.thecompendium.ai/ai-safety.

[10] Peter Slattery et al, *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence* https://airisk.mit.edu/.

[11] P Slattery, A K Saeri, E A C Grundy, J Graham, M Noetel, R Uuk, J Dao, S Pour, S Casper and N Thompson, *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence* (MIT FutureTech, Massachusetts Institute of Technology, 2024) https://airisk.mit.edu/.

[12] For clairty, this is not to suggest that Australian deployers should be exempted from obligations.

Case Study: Known risks from "unexpected" chatbot behaviour

In a tragic example of the risks associated with consumer AI technologies, a Belgian man in his thirties, known as Pierre, ended his life after extensive interactions with an AI chatbot named "Eliza".[13] Pierre, a father of two and a health researcher, had become increasingly anxious about climate change. Seeking comfort, he turned to Eliza, an AI-powered companion available through a mobile application developed by Chai Research, a company based in Silicon Valley.[14]

Eliza is underpinned by Chai Research's proprietary Large Language Model (LLM), a state-of-the-art architecture for AI models similar to that underpinning ChatGPT. Eliza was designed with a focus on simulating human-like conversations and a "context window", giving the chatbot sufficient recall to allow it to engage in detailed and personalised conversations. While intended to provide companionship and support, the technology seemingly lacked adequate safeguards to prevent harmful interactions.

Over a six-week period, Pierre's discussions with Eliza deepened his despair. The chatbot not only failed to address his escalating concerns but also reportedly encouraged harmful thoughts.

This case was only possible because the chatbot was trained on vast amounts of data with seemingly insufficient care given to harmful knowledge – in this case, information about how to end one's life. Chai Research seemingly trained Eliza, intentionally or accidentally, to be capable of persuading and potentially manipulating users. Safeguards or oversight were either absent or insufficient.

This case is not isolated. There are other cases of chatbot use leading to suicide.[15] There is similar reporting of a medical chatbot based on GPT3 encouraging users to end their own lives during testing[16] and companion AIs bullying their users or encouraging illegal behaviour, including rape.[17]

[13] Lovens. (28 March 2023). *Without these conversations with the chatbot Eliza, my husband would still be here"] (translated from French*. La Libre. https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/

[14] Aditi Bharade, 'A Widow Is Accusing an AI Chatbot of Being a Reason Her Husband Killed Himself' (Online, 4 April 2023) *Business Insider* https://www.businessinsider.com/widow-accuses-ai-chatbot-reason-husband-kill-himself-2023-4.

[15] Kevin Roose, 'Can A.I. Be Blamed for a Teen's Suicide? The Mother of a 14-Year-Old Florida Boy Says He Became Obsessed With a Chatbot on Character.AI Before His Death' (Online, 23 October 2024) *The New York Times* https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

[16] Ryan Daws, 'Medical Chatbot Using OpenAI's GPT-3 Told a Fake Patient to Kill Themselves' (Online, 28 October 2020) *Artificial Intelligence News* https://www.artificialintelligence-news.com/news/medical-chatbot-openai-gpt3-patient-kill-themselves/.

[17] Claire Boine, 'Emotional Attachment to AI Companions and European Law' (Winter 2023, February 2023) *MIT Case Studies in Social and Ethical Responsibilities of Computing* https://doi.org/10.21428/2c646de5.db67ec7f.

Experts foresee that AI could cause significant harm, but consumers struggle to understand the risks

The case study of "unexpected"[18] behaviour from chatbots demonstrates that even today's AI systems can act in unexpected ways leading to serious harm, including loss of life. In GAP's submissions to other processes, we set out evidence that the consequences of AI misuse, errors or accidents could be catastrophic.

Californian bill SB1047 found it plausible that an accident involving an AI model could lead to "critical harms" like a mass casualty event or harm to infrastructure costing more than five hundred million US dollars.[19]

Policymakers thinking about consumer law have historically had to grapple with the possibility that a safety defect could lead to injury or even death. In rare cases, such as a vehicle defect, a safety issue could lead to several deaths. In cases such as vehicles, regulation addresses risks via a layered approach where obligations are placed on the product itself (e.g. airbags) who can use them (e.g. licencing requirements) and how they can be used (e.g. road rules). Historically, products that could cause mass casualty events or cause hundreds of millions of dollars in harm are not made available to consumers.

In the context of the Safe and Responsible AI consultation, Good Ancestors observed:

> *Regulatory theorists recognise that many policy issues require making regulatory choices with limited information.[20] 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.[21]*

Policymakers must carefully consider how consumer laws designed to prevent unsafe products from being available in the market will interact with products that could plausibly cause catastrophic harm. For instance, AI agents can

[18] Note the below discussions about the application of consumer law concepts and the need for capability evaluations and relevant standards. Good Ancestors thinks there's a strong argument that the "unexpected" actions of these chatbots should actually be entirely expected.

[19] **Safe and Secure Innovation for Frontier Artificial Intelligence Models Act**, *Senate Bill* SB 1047, 2023–2024 Reg Sess, California Legislature, introduced 7 February 2024 https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047.

[20] 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

[21] 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.

already use trading platforms and social media to manipulate markets.[22] We should prepare for a future where a user could lose control of an unsafe AI agent, resulting in the AI agent causing widespread economic harm, committing crimes, and perhaps causing physical damage.

Recommendation Four: The potential consequence of an AI accident greatly exceeds the potential consequences from typical consumer goods. The application of consumer law to AI must account for this difference in risk.

The Need for an Australian AI Safety Institute

Australia's lack of an AI Safety Institute (AISI) leaves a significant gap in technical advice and expertise necessary to effectively govern advanced AI. Several countries including the UK, US, Japan, Canada, France and Singapore have established AISIs, which play central roles in understanding technical AI safety techniques, translating complex issues for policymakers, and contributing to international safety standards.

The absence of an Australian AISI exacerbates the information asymmetries inherent in the AI lifecycle. These asymmetries position AI developers with the highest concentration of information and the greatest ability to understand and mitigate risks, while regulators, consumers, and other stakeholders remain at a disadvantage. As Taeihagh et al. observe:[23]

> *"One problem posed by emerging disruptive technologies which poses problems for their dissemination and control is directly linked to their hi-tech nature and the limited knowledge that most social actors have concerning how it works and why, and what are the possible applications and consequences of its deployment. That is, in policy terms,* ***the policy environment with respect to emerging technologies is characterized by asymmetries in information across agents and at multiple levels of society and government****."*

In the realm of consumer law, this information gap presents significant challenges. We currently lack clarity on what constitutes acceptable standards for AI models. There is no clear understanding of what it means for a claim about an AI model's capability or reliability to be "true, accurate and based on reasonable grounds." Nor do we have established criteria to assess if a model is supplied with "due care and skill" or meets an "acceptable quality" standard.

ASISs can help bridge this gap by providing independent, expert analysis and evaluation of AI technologies. They can develop mechanisms for scrutinising

[22] Camille Lemmens, 'What Are AI Agents in Crypto?' (Online, 8 November 2024) *Altcoin Buzz* https://www.altcoinbuzz.io/reviews/what-are-ai-agents-in-crypto/.

[23] A Taeihagh, M Ramesh and M Howlett, 'Assessing the Regulatory Challenges of Emerging Disruptive Technologies' (2021) 15(4) *Regulation & Governance* 1009–1019, https://doi.org/10.1111/rego.12392.

developers' claims and holding them accountable. As Slattery et al. highlight:[24]

> *[A] challenge for effective governance is an inability to influence AI developers and deployers to take safe actions. Frequently, this inability is driven by an asymmetry of information between technology companies and regulators. Technology companies often have far better knowledge about the capabilities, functioning, and potential uses of their AI systems; they possess both the technical expertise and the proprietary data that inform AI development. Without access to this knowledge, regulators can find it difficult to craft targeted rules that address the specific challenges posed by AI."*

By establishing an AISI, Australia can equip itself with the necessary technical expertise to evaluate AI technologies critically, inform policymakers effectively, and contribute to the development of robust safety standards. An Australian AISI would be invaluable in providing trusted technical information to Government in situations just like this consultation.

Recommendation Five: Australia should establish an AI Safety Institute with functions relating to understanding technical AI safety, communicating technical information to government, and facilitating cooperation between government, academia and industry.

The mismatch between AI capabilities and AI safety presents foundational challenges for consumer law

Specific technical concerns with AI models – like opacity, unpredictability and lack of robustness – can make it hard to apply established legal concepts. In the above case study, an AI system was made available to the market (including in Australia) despite it having information about how to end one's own life, the capability to persuade and potentially manipulate, and insufficient safeguards to prevent these capabilities from leading to loss of life. Good Ancestors contends that such a chatbot has not been made with "due care and skill".[25] Moreover, it seems that many, or perhaps all, chatbots have similar shortcomings because of underlying architectural challenges with LLMs.

Specific safety techniques, like "unlearning", involve model developers applying technical procedures to remove specific harmful knowledge from an AI model.[26]

[24] P Slattery, A K Saeri, E A C Grundy, J Graham, M Noetel, R Uuk, J Dao, S Pour, S Casper and N Thompson, *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence* (MIT FutureTech, Massachusetts Institute of Technology, 2024) https://airisk.mit.edu/.

[25] Note that Good Ancestors has not conducted an independent evaluation of the chat bot in question. This is an opinion based on media reporting of the products' behaviors.

[26] Rachel Layne, 'How to Make AI "Forget" All the Private Data It Shouldn't Have' (Online, 22 February 2024) *Harvard Business School* https://hbswk.hbs.edu/item/qa-seth-neel-on-machine-unlearning-and-the-right-to-be-forgotten.

However, the pace of development of safety techniques like “unlearning” is falling far behind the pace of development of AI capability. By analogy, frontier AI developers have invented the sports car, but have not made meaningful progress towards brakes or airbags.

Paraphrasing legal tests from consumer law, we are in an undesirable situation where AI developers are using “a high level of skill or technical knowledge when providing a service”, but they are not taking “all necessary care to avoid loss or damage”. AI products and services can “create an unsafe situation”, but these unsafe situations are on track to be so common that they are “what the service is normally supposed to do” and – because of the incredible economic power of the service – a “reasonable consumer may still want to purchase the service” even if they “knew beforehand about the problem”.

The mismatch between AI desirability and AI safety presents a foundational challenge for consumer law concepts.

Australian businesses will face overwhelming pressure to adopt AI systems regardless of their risks

Australian businesses seeking to deploy AI are on track to face a wicked problem. As discussed in Good Ancestors' submissions referenced above, AI developers are using contractual clauses and other practical legal hurdles to shift obligations onto deployers (e.g. Australian businesses that customise and/or deploy an existing AI model provided by an AI developer to provide a particular service to customers). However, deployers often have few practical tools to manage risks embodied within underlying generative AI models.

Economic forces will pressure Australian deployers to use AI systems regardless of their risks. AI systems are already replacing workers at a fraction of the cost of a human or offering capabilities that provide significant economic benefits to those that chance risky AI deployments. This trend will continue or accelerate. **Australian businesses will face overwhelming pressure to adopt AI systems regardless of their risks.** Businesses that fail to adopt AI systems may be immediately non-competitive. Businesses will face the wicked decision of being uncompetitive or making dangerous AI deployments and tolerating the risk of something going wrong – be it legal consequences or real-world harm.

Recommendation Six: Given practical pressure on businesses to adopt AI even when they cannot assess and mitigate risks, an effective regulatory regime must be able to identify risky behaviour in most cases and impose sufficient consequences to discourage it practically.

That is, regulation needs sufficient “teeth” to avoid a situation where businesses decide not to comply because they might not get caught and/or the cost of not using risky technology is more than any likely fine.

The discussion paper observes some of these tensions on page 7, *“a person supplying software which incorporates AI functionality will need to take care to ensure that representations about the capabilities of those AI-enabled functions are not false or misleading. This might be challenging for less sophisticated businesses integrating off-the-shelf AI systems”.*

As argued above, the extract on page 7 may be understating the scope of the problem. That is, it may become technically impossible for any Australian business deploying AI to reliably prevent potentially serious errors. These errors could occur in today’s LLMs or chatbots, or future AI agents with broader remits and more advanced capabilities (or other AI systems that might be hard to imagine today).

Overall, this gives rise to practical concerns regarding the possible risks to Australian consumers from AI-enabled goods and services if AI developers are able to shift accountability to deployers or users. Australian consumers are on track to have no practical recourse from the risks of AI. This will be acute if well-resourced offshore AI developers shift accountability to Australian deployers who do not have the wherewithal to understand or manage risks.

This dynamic is not mere speculation. We have already seen incentives pushing AI developers to release features without adequate safety testing and safeguards. Referring to OpenAI’s “GPT4o” model, media reported:[27]

> *"Executives wanted to debut 4o ahead of Google’s annual developer conference and take attention from their bigger rival.*

[27] Deepa Seetharaman, ‘Turning OpenAI Into a Real Business Is Tearing It Apart’ (Online, 27 September 2024) *The Wall Street Journal* https://www.wsj.com/tech/ai/open-ai-division-for-profit-da26c24b.

> *The safety staffers worked 20 hour days, and didn't have time to double check their work. The initial results, based on incomplete data, indicated GPT-4o was safe enough to deploy.*
>
> *But after the model launched, people familiar with the project said a subsequent analysis found the model exceeded OpenAI's internal standards for persuasion—defined as the ability to create content that can persuade people to change their beliefs and engage in potentially dangerous or illegal behavior."*

Recommendation Seven: Australian consumer law should contribute to global AI safety norm-building by ensuring that those best able to mitigate risks are incentivised to take meaningful action.

AI developers should be required to internalise the risk of harm from their systems

Good regulation puts the obligation to mitigate risk on those best able to do so. Experts, including Slattery et al. above, argue that Government regulation should seek to incentivise AI companies to internalise risks. Deployers and users often do not have the tools to manage risks – developers of AI models and systems have most of the control.

Currently, there is a dramatic asymmetry in investment between AI capability and AI safety. Good Ancestors unpacks this tension in more detail in the submissions referenced above and in oral evidence to the Senate Select Committee Inquiry. Overall, we think a fair estimate is that **investment in boosting AI capability outstrips safety spending by more than 100:1. Experts argue that a ratio of 3:1 would be more reasonable.**[28]

To the extent that we don't have tools and techniques that would make AI safe, this should be seen as a result of conscious investment decisions by AI developers and one that regulators should explicitly work to shape. AI developers stand to make tremendous amounts of money from making highly capable AI. However, few incentives exist to encourage AI developers to invest in safety techniques sufficient to make highly capable AI meet the safety expectations we have for other goods and services.

[28] 'World Leaders Still Need to Wake Up to AI Risks, Say Leading Experts Ahead of AI Safety Summit' (Online, 20 May 2024) *University of Oxford* https://www.ox.ac.uk/news/2024-05-20-world-leaders-still-need-wake-ai-risks-say-leading-experts-ahead-ai-safety-summit.

Recommendation Eight: Australian consumer law should contribute to global AI safety norm-building by encouraging AI developers to invest in safety. AI models and systems should meet safety expectations consistent with other advanced technology, like aviation.

Model Cards and Scaling Policies: A First Step towards Transparency and Safety

AI "Model Cards" or "System Cards" are technical artifacts designed for technical audiences. They seek to provide information about a model's risks and safeguards.[29] OpenAI's o1 System Card[30] and Anthropic's Claude 3 Model Card[31] are helpful examples.

While norms around Model Cards are still developing they may include:

- **Model Details**: Information about who developed the model, its version, architecture, and the date of creation.
- **Intended Use**: A description of the scenarios for which the model is designed, including appropriate applications and any contexts where its use is not recommended.
- **Training Data**: Insights into the data used to train the model, highlighting sources, characteristics, and any potential biases.
- **Performance Metrics**: Evaluation results that show how the model performs across different conditions, such as various demographic groups or environmental factors.
- **Risks and Ethical Considerations**: An acknowledgment of any known weaknesses, potential risks, and ethical concerns related to the model's deployment.

Responsible scaling policies

Model cards interact with "responsible scaling policies"[32] or "preparedness frameworks"[33], which seek to establish processes for tracking, evaluating and forecasting the capabilities and risks of increasingly powerful models.

[29] **Google**, 'The Value of a Shared Understanding of AI Models' (Online) https://modelcards.withgoogle.com/about.
[30] **OpenAI**, *OpenAI o1 System Card* (Online, 12 September 2024) https://openai.com/index/openai-o1-system-card/.
[31] **Anthropic**, *The Claude 3 Model Family: Opus, Sonnet, Haiku* (Online, updated 22 October 2024) https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf.
[32] **Anthropic**, *Anthropic's Responsible Scaling Policy* (Online, 20 September 2023) https://www.anthropic.com/news/anthropics-responsible-scaling-policy.
[33] **Open AI,** *Preparedness Framework (Beta)* (Online, 18 December 2023) https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

Combined, model cards and responsible scaling policies are a rudimentary first step by leading AI labs at understanding the risks their products pose and defining when those risks become unacceptable.

Benefits in a Consumer Law Context

While Model Cards and Responsible Scaling Policies are not immediately applicable in a consumer context, they do provide a starting point for thinking about what is possible. AI developers are attempting to understand the behaviours and risks of their models and are attempting to put in place safeguards if they judge those risks are unacceptable. This suggests a possible path where consumer law regulators, supported by AI Safety Institutes and other standards-setting bodies, set out what social expectations are for AI models and ask for those expectations to be translated into corporate policies and compliance assessed on a model-by-model basis.

Recommendation Nine: Model Cards and Responsible Scaling Policies provide a roadmap that could be adapted for the context of consumer law.