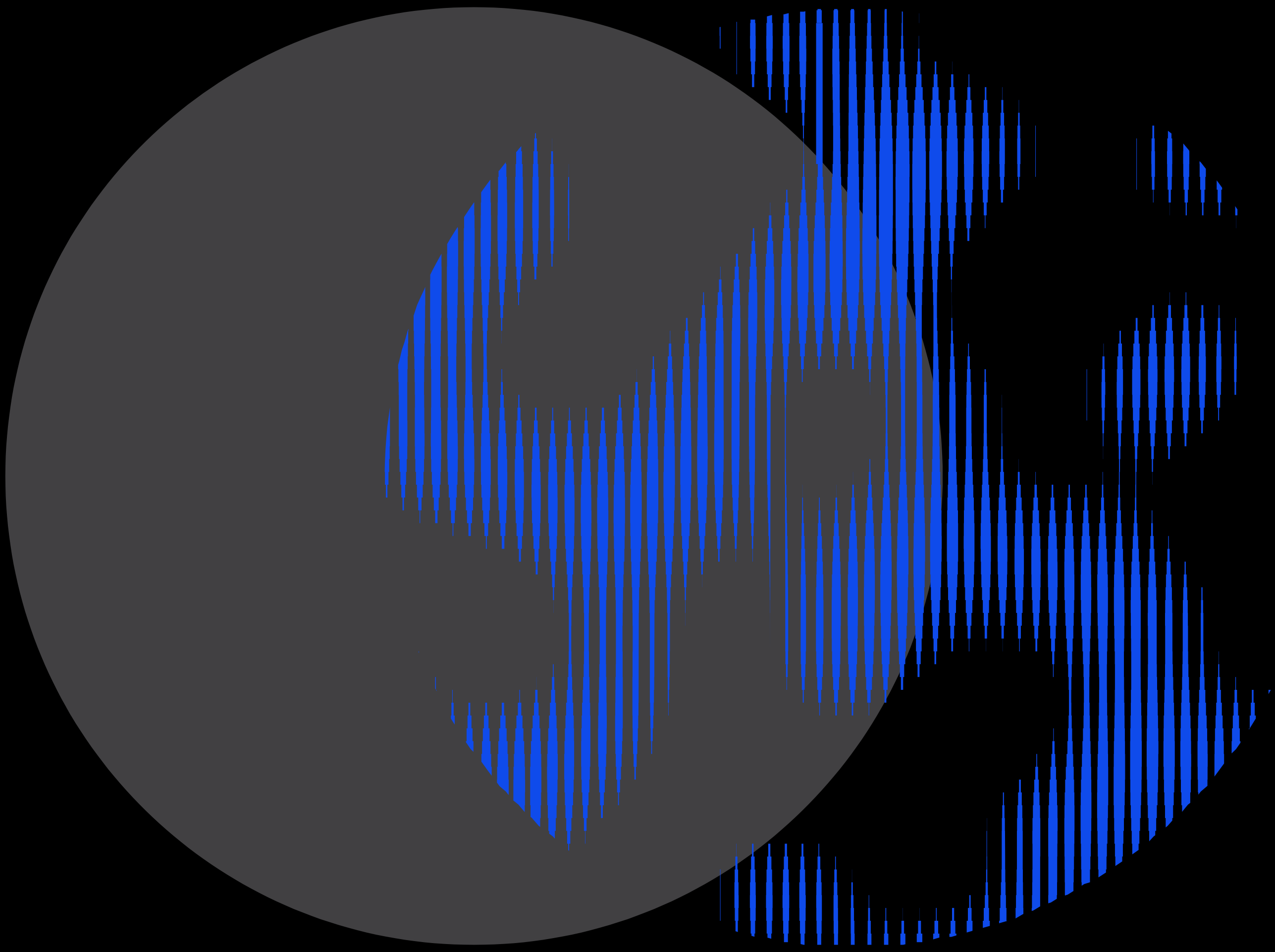
Human Technology
Institute



# Safe and Responsible AI in Australia: Proposals Paper for introducing mandatory guardrails in high-risk settings

*Submission to the Department of Industry, Science and Resources*

4 October 2024

4 October 2024

The Human Technology Institute (HTI) is building a future that applies human values to new technology. HTI embodies the strategic vision of the University of Technology Sydney (UTS) to be a leading public university of technology, recognised for its global impact specifically in the responsible development, use and regulation of technology. HTI is an authoritative voice in Australia and internationally on human-centred technology. HTI works with communities and organisations to develop skills, tools and policy that ensure new and emerging technologies are safe, fair and inclusive and do not replicate and entrench existing inequalities.

The work of HTI is informed by a multi-disciplinary approach with expertise in data science, law and governance, policy and human rights.

HTI provides independent advice and other input to the Australian Government, as well as state and territory governments, on AI reform. Professor Edward Santow, HTI's Co-Director, is a member of the temporary AI Expert Group set up by the Department of Industry, Science and Resources (DISR) to advise Government on high-risk AI.

**For more information, contact us at hti@uts.edu.au**

**Authors**: Sophie Farthing, Lauren Perry, Sarah Sacher and Professor Edward Santow

### Acknowledgement of Country

UTS acknowledges the Gadigal people of the Eora Nation, the Boorooberongal people of the Dharug Nation, the Bidiagal people and the Gamaygal people upon whose ancestral lands our university stands. We would also like to pay respect to the Elders both past and present, acknowledging them as the traditional custodians of knowledge for these lands.

# Contents

# Tables

4 October 2024

# Executive summary

The Australian Government has committed to introducing legislation for artificial intelligence (AI) in a way that 'builds community trust and promotes innovation and adoption while balancing critical social and economic policy goals'.[1] The Government has said that its regulatory response to AI will be balanced and proportionate, risk based, and put people and communities at the centre.[2]

The Australian Government is undertaking a range of reform processes. Many of those processes are considering reform to certain substantive areas of law where AI has particularly significant implications – such as privacy law and copyright law. Such reform is motivated, at least in significant part, by the rise of AI. In other words, those laws are intended to be *informed by* the operation of new and emerging technologies such as AI. However, those laws are primarily expressed in *technology-neutral language*. This reflects the conventional regulatory approach in Australia and in the vast majority of jurisdictions: it is rare for the principal regulatory object of a law to be AI or any other technology for that matter. The vast majority of laws set requirements that apply to all technologies and none.

While most of Australian legislation is technology neutral, this is only the default position. Some of our laws are technology specific, in that they are directed explicitly to one or more technologies. The Government's *Safe and Responsible AI in Australia* Proposals Paper (the Proposals Paper) considers technology-specific, as distinct from technology-neutral, law reform. We understand that the reform contemplated by the Proposals Paper is intended to complement other reform that the Government is undertaking, in areas as diverse as copyright, automated decision making and digital platforms.

The Proposals Paper outlines options to introduce legislation that would create mandatory guardrails for high-risk AI. Those mandatory guardrails are, in essence, requirements that AI developers and deployers would be required to follow as they develop and deploy AI models and systems that fall within a statutory definition of 'high-risk AI'. In this way, the mandatory guardrails legislation proposed in the Proposals Paper could provide an economy-wide approach that operates in tandem with other reform that responds to the rise of AI.

The Human Technology Institute (HTI) does not express a view in this submission on the desirability or otherwise of the Government's overall mix of technology-neutral and technology-specific law reform. Instead, HTI assumes that the Government is minded to take forward reform along the lines of that set out in the Proposals Paper, and this submission offers our views on how this potential reform could be refined or improved. It is particularly important that the Government adopt an effective legislative model for the mandatory guardrails that will incentivise responsible design and development; clarify who is accountable, in law, for compliance with the guardrails; and build on current laws to ensure accessible avenues of redress to enforce compliance with the mandatory guardrails.

This submission is divided into three parts as follows:

## Part A: defining high-risk AI

HTI observes that the regulatory object of the Government's proposed mandatory guardrails legislation is high-risk AI. As 'high-risk AI' is not a term of art with a generally understood meaning, the Government's mandatory guardrails law will need to include **a clear statutory definition of each of the two limbs in the term 'high-risk AI'**. For the first limb, HTI proposes **a conventional, matrix approach to determining the level of risk posed.** This would involve weighing of a number of interrelated factors that, in aggregate, support an overall assessment of whether or not the AI model or system presents a high risk. HTI proposes that **the legal foundation for this risk assessment be Australia's obligations under international human rights law**. HTI recommends that the second limb of this term **(the meaning of 'AI') should be defined by reference to a statutory definition derived from another leading jurisdiction**, such as the European Union (EU) AI Act.

The definitional approach recommended by HTI **most closely resembles the principles-based approach to defining risk** outlined in the Proposals Paper. HTI recommends a number of amendments to enhance the draft principles (or factors) by providing greater certainty and specificity.

While AI developers and deployers will assess for themselves whether their AI models and systems fall within the scope of 'high-risk AI', HTI also recommends that the Government consider **measures that would support greater certainty in these assessments**. This would include, at a basic level, providing authoritative guidance on understanding and applying the factors (or principles) in the definition. It may also include a mechanism for official review and certification of assessments of AI developers and deployers.

**HTI cautions against broad exemptions** to the proposed mandatory guardrails law for all defence and national security bodies regardless of their activities. Any exceptions to the application of this law should be carefully circumscribed to apply only where it is demonstrably justified by reference to international human rights law.

HTI recognises that **some AI models, systems and technologies can pose unacceptable risks to human rights**. There should also be a process for the relevant Minister to determine that certain exceptional models, systems or technologies fall within the category or high or unacceptable risk, where it would represent an unjustifiable limitation of human rights.

## Part B: Guardrails ensuring testing, transparency and accountability of AI

In this part of the submission, **HTI proposes some refinements to the mandatory guardrails** as they were outlined in the Proposals Paper, including clarifying the requirements and responsibilities of developers and deployers, to improve risk mitigation of high-risk AI.

We also propose consideration of **two new guardrails**. First, in specified contexts, AI developers and deployers should be required to **engage with stakeholders** to evaluate their needs and circumstances as part of the safe development and deployment of AI models and systems. Secondly, all organisations should be required to have a plan for the **safe decommissioning** of high-risk AI systems. HTI also considers how the guardrails may better safeguard First Nations peoples, knowledge and cultural protocols; and how small-to-medium sized enterprises can be supported to apply the guardrails.

## Part C: Regulatory options to mandate guardrails

The Proposals Paper outlines three broad legislative models for incorporating the mandatory guardrails in law. HTI considers that, **with careful drafting, either of Options 2 or 3 would be suitable** to introduce the mandatory guardrails.

In this part of the submission, HTI makes a number of recommendations regarding the substantive content of the Government’s proposed legislation. That is, the proposed legislation should:

- set out a **clear and unambiguous objective** to protect people from harm, and to support innovation for economic benefit and societal wellbeing
- require AI developers and deployers to take **reasonable steps to comply** with the mandatory guardrails
- contain a **rebuttable presumption** that where a person is responsible for making a decision using AI, that person is legally liable for the impact of that decision
- **provide for enforcement** through appropriate mechanisms such as oversight by a regulator, and a ‘piggy back’ provision that would support people with an existing cause of action based on s 39(1) of the *Charter of Human Rights and Responsibilities 2006* (Vic).

HTI recommends the Government **incentivise compliance** with the mandatory guardrails through measures such as practical, sector-specific advice or guidance; preference in government procurement for those companies that can demonstrate compliance with the mandatory guardrails; and a mechanism to reduce the liability of AI developers and deployers that have demonstrated good-faith compliance with the mandatory safeguards.

# List of recommendations

### Recommendation 1

HTI recommends that the proposed mandatory guardrails legislation define 'high-risk AI' by reference to a matrix of principles or factors that include:

a. use contexts that are generally high risk

b. the likely impact on one or more individual's human rights, as that term is defined in s 3(1) of the *Human Rights (Parliamentary Scrutiny) Act 2011* (Cth)

c. the risk of adverse impacts to an individual's physical or mental health, their safety and risk of financial loss or property damage

d. the risk that the AI system will have a legal or similarly significant effect on an individual

e. the risk of adverse impacts to groups of individuals or collective rights of cultural groups

f. the risk of adverse impacts to the broader Australian economy, society, environment, rule of law or liberal-democratic system

g. the upside risk or likelihood to bring relevant public benefits

h. the severity, extent and relative likelihood of any relevant adverse or positive impact.

### Recommendation 2

HTI recommends the mandatory guardrails legislation include a mechanism for an AI developer or deployer to seek review, by an independent authority such as a regulator, of its own assessment regarding whether an AI system or model falls within the definition of 'high-risk AI'.

### Recommendation 3

HTI recommends that the proposed mandatory guardrails legislation not contain a broad exemption for defence and national security organisations. Where a defence or national security organisation considers that this proposed law should be subject to an exception, there should be an independent adjudicative process to consider an application and decide whether the application should be granted, and on what terms.

### Recommendation 4

HTI recommends that the proposed mandatory guardrails law enable the relevant Minister to assess whether certain AI systems or technology pose an unjustified restriction on human rights. In that scenario, the Minister should be able to designate by legislative instrument the relevant AI system or technology as high risk or prohibited.

### Recommendation 5

HTI recommends a number of refinements to the current draft of the mandatory guardrails (presented in Table 2), including further detail and clarification of the responsibilities between developers and deployers.

### Recommendation 6

In addition to the refinements to the Proposals Paper's current list of mandatory guardrails, HTI recommends that the Government consider the following additional mandatory guardrails:

a. a requirement to develop a plan for the safe decommissioning of high-risk AI models or systems.

b. a requirement for government to engage with stakeholders and evaluate their needs and circumstances in the development and deployment of high-risk AI models or systems

c. a requirement on the private sector to undertake stakeholder consultation in respect of high-risk AI systems, either via an additional mandatory guardrail or by expanding one or more of the existing proposed guardrails.

### Recommendation 7

HTI recommends that the Government reduce the risk of AI systems harms to First Nations peoples, languages, cultures and knowledge by enacting a mandatory guardrail for consultation in particular high-risk contexts (per Recommendation 6).

### Recommendation 8

HTI recommends that the Government provide clear, practical guidance to organisations on how to comply with the mandatory guardrails – including targeted resources and support for SMEs. This support should come from an appropriately resourced regulator or peak body.

### Recommendation 9

If the Government introduces legislation containing a list of mandatory guardrails for developers and deployers of high-risk AI, the legislation should:

a. include a definition of 'high-risk AI', in the manner set out in Recommendation 1

b. set out a clear regulatory objective

c. require AI developers and deployers to take reasonable steps to comply with the mandatory guardrails

d. contain a rebuttable presumption that where a person is responsible for making a decision using AI, that person is legally liable for the impact of that decision

e. provide for enforcement through appropriate mechanisms such as oversight by a regulator, and a 'piggy-back' provision that would support people with an existing cause of action based on s 39(1) of the *Charter of Human Rights and Responsibilities 2006* (Vic).

**Recommendation 10**

HTI recommends the Government consider additional measures to support compliance with the proposed mandatory guardrails legislation, including:

a. guidance from a regulator or other authoritative body

b. amending procurement rules to prioritise companies demonstrating compliance with the mandatory guardrails

c. reducing, to an appropriate extent, the relevant legal liability of an organisation that has demonstrated compliance with the mandatory safeguards.

# Part A: Defining high-risk AI

**Question 1:** Do the proposed principles adequately capture high-risk AI? Are there any principles we should add or remove?

Please identify any:

- low-risk use cases that are unintentionally captured
- categories of uses that should be treated separately, such as uses for defence or national security purposes.

**Question 3:** Do the proposed principles, supported by examples, give enough clarity and certainty on high-risk AI settings and high-risk AI models? Is a more defined approach, with a list of illustrative uses, needed?

- If you prefer a list-based approach (similar to the EU and Canada), what use cases should we include? How can this list capture emerging uses of AI?
- If you prefer a principles-based approach, what should we address in guidance to give the greatest clarity?

## Approach to risk

### Need for a legally certain articulation of risk

The Proposals Paper adopts a risk-based approach and sets out draft principles for determining whether an AI system is high risk.

A risk-based approach to AI is consistent with the reform approach of many other leading jurisdictions, as well as international bodies such as the OECD and UN AI Advisory Body.[3] However, there are many views on precisely *what a risk-based approach means*. Risk in the AI regulatory context is generally understood to focus on a general category of downside risk – namely, ‘risk of harm’ to people. Yet this is not a term of art. It is unclear what specific harms are within the ambit of this term, and there is a lack of consensus regarding how harms should be categorised and weighted.

As the proposed mandatory guardrails would apply to AI systems that are considered ‘high risk’, the law will need to provide a definition of risk. Parliament could choose to adopt an entirely novel definition, but that would make compliance more difficult, especially for AI developers and deployers that operate across multiple jurisdictions. Hence, HTI proposes that the terms ‘risk’ and ‘risk of harm’ be defined by reference to an established legal norm. This would promote a consistent approach to risk across government, developers, deployers and the general public, including to guide risk assessment processes, and the interpretation and enforcement of AI-related laws.

### International human rights law as a normative foundation

HTI recommends that the definitions of ‘risk’ and ‘harm’ should be grounded in Australia’s obligations under international human rights law. While there is no federal

Human Rights Act, the Australian Parliament has set out the primary human rights recognised in Australian law in the *Human Rights (Parliamentary Scrutiny) Act 2011* (Cth).[4] This list could be incorporated as the normative foundation, providing certainty for industry and the community about which harms are within scope of the mandatory guardrails legislation.

The Proposals Paper identifies human rights law as relevant to identifying AI risks, and it is implicit in the draft principles for defining high-risk AI that the Government is prioritising addressing harms to people in the development of AI laws. The principles themselves are broadly aligned with a human rights-based approach, but this is not explicit. Going a step further, articulating more precisely how risk should be interpreted would make the principles more coherent and straightforward to apply. It would have the effect of embedding a normative basis for understanding risk and a framework for weighing risk alongside various competing rights and interests.

Under international human rights law, risk would primarily be understood to refer to a range of harms (such as infringements of the right to privacy, the right to equality and the right to health) that are enumerated in international treaties, including the International Covenant on Civil and Political Rights (ICCPR), to which Australia is a party and which have been partially incorporated into domestic Australian legislation.

There are several benefits to adopting international human rights law as the normative basis for a definition of risk of harm:

- The Australian Government is already required under international law to consider harms by reference to human rights. Human rights are embedded in existing Australian law and policy. For example, they *must* be considered in the drafting of all primary and delegated legislation. Human rights compliance is also an important objective in a range of AI-related policy, including the Australian Government's *AI Ethics Principles*.
- International human rights law exists to prevent harm to humans. It sets out harms recognised by law, with clear definitions of the relevant harms. Those definitions have been the subject of extensive judicial and other authoritative consideration over many decades.
- International human rights law recognises that other legitimate interests, including commercial and economic imperatives, should be given due weight, including in any risk assessment. It provides an effective mechanism for addressing multiple risks simultaneously. Where those risks come into tension with each other, it provides a mechanism for reconciling those risks, especially via the proportionality test. It recognises that not all harms are of equal severity, and so focuses attention on the most serious harms.
- International human rights law sets out distinct responsibilities for business and government. The UN Guiding Principles on Business and Human Rights authoritatively state how businesses should fulfil their own responsibilities across the value chain. International human rights law also contains specific responsibilities for government, which both encourages government to act as an exemplar when developing and using AI and provides added protections for individuals in the context of government decision making.

- An approach to risk grounded in human rights would support the Government's stated goal of enabling interoperability with other leading jurisdictions, such as the EU, UK, Canada and the United States, which all take a rights-based approach. Adopting human rights law would also ensure Australians are protected to the same standard as those in comparable liberal democracies.
- A human rights-based approach also ensures that people most at risk of AI harms are kept at the forefront of responses to AI. AI harms are not dispersed or experienced equally by all. AI can disproportionately harm certain groups, including women, children, people with disability, LGBTQIA+ people, older people, Aboriginal and Torres Strait Islander people, and those from culturally and linguistically diverse backgrounds.

# Assessing risk

## Need for a risk matrix

A risk-based approach to AI should enable consideration of all relevant risks, including economic, commercial, social and political risks. Human rights law enables these risks to be given due weight, and where risks come into tension with each other, provides a mechanism for reconciling those risks.

The draft principles should be amended to provide greater clarity about how risk should be classified and weighed in practice. A human rights-based approach would support the development of a risk matrix that takes into account a range of risks and mitigating factors, such as:

- upside risks (ie, positive opportunities) and downside risks (ie, threats)
- a broad range of risks that government and the private sector typically consider would be relevant—including economic, commercial, social, political, environmental and safety risks
- the context in which the relevant risk arises. For example, AI used in high-stakes decision-making contexts, such as law enforcement, would involve a higher level of risk than, say, AI in a computer game
- any risk mitigants, including human oversight and governance safeguards, as well as the existence of effective regulation in respect of a particular area or activity
- a proportionality-based mechanism for balancing risks, rights and interests that come into conflict with each other.

A risk matrix should be included in the mandatory guardrails legislation.

**Box 1: Weighing risks, interests and rights under international human rights law**

International law recognises that human rights protections may need to be balanced against other human rights and legitimate interests. The Siracusa Principles state that non-absolute human rights may be subject to limitations only where the limitations are lawful, and can be demonstrably justified in a free and democratic society.

The following factors are considered when determining whether a limitation of a right is reasonable and justified:

- whether the limitation is in pursuit of a legitimate purpose (this encompasses a range of lawful purposes).
- whether the limitation has a rational connection to the purpose to be achieved. This requires a reasoned and evidence-based explanation as to how the proposed approach is likely to be effective in achieving the legitimate purpose.
- whether the limitation on rights is necessary to achieve the legitimate purpose. This includes consideration of:
    - whether there are any reasonably available means to achieve the purpose that are less restrictive of human rights
    - the extent of the interference with the human right (the greater the interference, the less likely it will be proportionate)
    - whether there are safeguards or controls over the measures adopted (for example, oversight measures or avenues for redress)
    - whether affected groups are particularly at-risk or may experience disproportionate impacts.[5]

## ‘Principles based’ or ‘list based’?

The Proposals Paper distinguishes between a ‘principles-based’ and a ‘list-based’ approach to defining risk.

HTI understands that a ‘list-based approach’, as outlined in the Proposals Paper, would set the level of risk primarily or exclusively based on the activity or context of the use of AI – for example, AI used in the health sector would always be deemed to be high risk. While it would be possible to add *some* nuance to a list-based approach by drafting the list with a large range of exceptions and other considerations, this would reduce the one major attraction of the list-based approach – namely, its conceptual simplicity.

By contrast, a principles-based approach would allow consideration of a range of factors, including those on the Proposals Paper list, such as use context. However, it would also allow consideration of other factors. For example, if an AI system were used for medical diagnosis, one factor – namely, that the AI system is being deployed in a healthcare context – would militate in favour of an assessment of high risk. However, if the AI system were a purely administrative tool, with no or minimal likelihood of harm to people, this would be a countervailing factor against an overall assessment of high risk. This example is described further in Box 2.

In other words, a list of use contexts, which suggest but do not definitively determine that the risk is high, could be incorporated into a principles-based approach. Whether or not the AI system in question is high risk would depend on the use context, as well as other factors that could increase or lower the level of risk for that specific AI system. HTI supports a principles-based approach that takes into account these use contexts as one of a number of possible indicators of high-risk AI. The level of risk posed by particular AI systems would be determined through the application of a risk matrix. HTI endorses the use contexts set out in the Proposals Paper which suggest a high risk, such as in law enforcement, critical infrastructure and employment. HTI also recommends that administrative decision making by government be added to this list.

HTI cautions against the adoption of *only* a list-based approach to risk. There are two particular problems with that approach. First, risk in many complex AI deployment contexts is multifaceted. A list-based approach focuses only on one facet, which makes it impossible to consider other factors relevant to a true or accurate assessment of risk. It could result in AI systems, which have a low or no likelihood of causing harm, being assessed as high risk simply because of their deployment context. It could also result in some AI systems that do carry a high likelihood of harm being assessed as low risk, simply because the deployment context generally appears to be low stakes.

The second problem with a list-based approach is that, while it appears to offer greater certainty than a principles-based approach, there is likely to be fierce debate about the parameters of the various categories on the list. For example, this could incentivise organisations to make false claims about the deployment contexts in which their AI systems are to be deployed.

In any case, there are other options for achieving greater certainty through a principles-based approach. This submission outlines how this can be achieved, including through HTI's proposed amendments to the principles, as outlined in Table 1.

**Box 2: Challenges of a list-based approach to classifying high risk**

Imagine an AI tool that manages the efficient use of stationery by hospital staff and has no direct bearing on access to healthcare or treatment of patients. It is likely to present only a low risk of harm to people. Yet, if all deployments of AI within healthcare were considered to be high risk, this AI tool would also be categorised as high risk, causing unnecessary compliance burdens.

Now, imagine a different AI tool used in a supermarket to automatically match the faces of customers to social media profiles in order to build a database of people's shopping behaviours without their permission. While supermarkets may seem like a lower risk context than hospitals, the human rights impacts of this tool mean it would likely be considered high risk.

## Mandatory guardrails legislation should define AI and include other relevant definitions

As we have explained in other forums, HTI considers that the primary regulatory response to the rise of AI should be via *technology-neutral legislation*. This is a useful approach because technology-neutral laws are more adaptable to technological developments, better accommodating future innovations and associated risks. For this reason, technological neutrality is the conventional and most common approach that Australian law has applied to almost all technologies. That this is the desirable *primary* response acknowledges a role – albeit a secondary role – for legislation that makes AI in general, or certain types of AI, the regulatory object.

Without expressing a view on whether technology-neutral law reform should be prioritised more highly than the Government has chosen to date, it suffices to observe that this Proposals Paper is premised on adopting legislation that is *technology specific*, as opposed to *technology neutral*. In other words, the Proposals Paper makes clear that the regulatory object for the proposed mandatory guardrails is, explicitly, 'high-risk AI'.

In addition to defining the term 'high risk', it will also be necessary to define the term 'AI'. There is no universally accepted legal definition of AI, and this concept is notoriously difficult to define with legal precision. Therefore, to promote internationally interoperable laws in this area, the Government should consider adopting a definition that has legislative force in another major jurisdiction, such as the EU's AI Act.[6] For example, under Article 3 of the AI Act, 'AI system' is defined to mean:

> a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[7]

## Commentary on wording of draft principles

As summarised below, HTI proposes some changes to the draft principles for defining high-risk AI.

**Table 1: Proposed changes to the draft principles for defining high-risk AI**

| Principle | HTI input |
|---|---|
| (a) The risk of adverse impacts to an individual's rights recognised in Australian human rights law without justification, in addition to Australia's international human rights law obligations | HTI supports consideration of the impact on people's human rights.<br>While the Proposals Paper refers to discrimination as a potential human rights violation, other human rights also need to be considered as part of a risk assessment process. These include, for example, freedom of expression; freedom of association; right to privacy; right to health; right to access essential services, such as social security; right to a fair trial; and protections against cruel, inhuman and degrading treatment. It would be helpful to make clear that this broader range of human rights is included within the scope of this term.<br>The wording of the principle implies that *all* human rights can be justifiably limited. This is not the case; some human rights are |

| Principle | HTI input |
|---|---|
| | absolute. These include freedom from slavery, freedom from torture, and the right to life.<br>The proportionality test is relevant to an assessment of *all* the principles, not just the first principle. It should assist in balancing all relevant risk considerations or factors. |
| (b) The risk of adverse impacts to an individual's physical or mental health or safety | The terms 'health' and 'safety' should be defined to provide certainty. In addition to including physical and mental harms, this principle should include financial loss and property damage. |
| (c) The risk of adverse legal effects, defamation or similarly significant effects on an individual | HTI supports this principle for the reasons outlined in the Proposals Paper, but suggests that it be reworded as follows for clarity: the risk of a 'legal or similarly significant effect on an individual'. This wording would align Australian law with EU law, such as the GDPR, supporting interoperability. The Australian Government also committed to adopting this formulation in the Privacy Act, in its response to the Privacy Review Report.[8]<br>It is unclear why defamation is singled out as the sole example of a legal right in this principle. It would be helpful to provide further indicative examples of legal effects, such as by reference to consumer rights, employment rights, social security entitlements, and contractual rights.<br>Indicative examples should also be provided to illustrate the meaning of 'significant effects.' GDPR guidance states significant effects are not necessarily legal effects, though 'the decision must have the potential to significantly influence the circumstances, behaviour, or choices of the individuals concerned'. 'Significant effects' may refer to, for example, decisions that significantly impact someone's financial circumstances (eg, automatic refusal of credit eligibility), employment opportunities (eg, e-recruitment).[9] |
| (d) The risk of adverse impacts to groups of individuals or collective rights of cultural groups | This principle rightly reflects AI's potential to have a disproportionate negative effect on disadvantaged, minority, marginalised or vulnerable groups. There are examples of this phenomenon being experienced by people with characteristics protected by anti-discrimination law, such as First Nations peoples, people with disability, children and other groups.<br>There are also examples of other groups being affected, such as those who rely on essential government services, or current and former prisoners. For instance, a CHOICE investigation revealed that an algorithm used by Airbnb was arbitrarily removing people |

| Principle | HTI input |
|---|---|
| | deemed to be sex workers from its platform, with no transparency, explanation or options for review.[10] |
| (e) The risk of adverse impacts to the broader Australian economy, society, environment and rule of law | This principle should be expanded to refer also to adverse impacts on Australia's liberal democratic system, such as AI that can be used to manipulate the conduct of democratic elections, or the proliferation of facial recognition systems which lead to unjustified or mass-surveillance. |
| The severity and extent of those adverse impacts outlined in principles (a) to (e) above. | In considering this principle, it would be necessary also to consider the relative likelihood of the relevant risk materialising, and the impact of any risk mitigants, such as human oversight. |

**Recommendation 1**

HTI recommends that the proposed mandatory guardrails legislation define 'high-risk AI' by reference to a matrix of principles or factors that include:

a. use contexts that are generally high risk

b. the likely impact on one or more individual's human rights, as that term is defined in s 3(1) of the *Human Rights (Parliamentary Scrutiny) Act 2011* (Cth)

c. the risk of adverse impacts to an individual's physical or mental health, their safety and risk of financial loss or property damage

d. the risk that the AI system will have a legal or similarly significant effect on an individual

e. the risk of adverse impacts to groups of individuals or collective rights of cultural groups

f. the risk of adverse impacts to the broader Australian economy, society, environment, rule of law or liberal-democratic system

g. the upside risk or likelihood to bring relevant public benefits

h. the severity, extent and relative likelihood of any relevant adverse or positive impact.

## Independent review of self-assessed risk level of an AI system or model

In the first instance, an AI developer or deployer should be required to make its own assessment of whether its AI models or systems fall within the definition of ‘high-risk AI’. However, where this self-assessment is incorrect, this would have the effect of either the organisation failing to comply with important mandatory guardrails, or incurring unnecessary expense in complying with guardrails that are inapplicable.

Even if all developers and deployers adopt a conscientious approach to this self-assessment, errors are likely, given that there is significant nuance and judgment necessary in weighing up the relevant factors in any categorisation of ‘high-risk AI’. Hence, there is a case for including in the mandatory guardrails legislation a mechanism for a regulator or other independent certifier to confirm or change the developer or deployer’s self-assessment. This would have the effect of offering the organisation certainty regarding whether it must comply with the mandatory guardrails. Applying for the review could be voluntary, or it could be compulsory in some circumstances or contexts (such as when AI is used in weapons).

This kind of certification or regulator oversight is not uncommon, especially in areas such as planning and environmental law. In Australia, the Australian Competition and Consumer Commission (ACCC) issues certification trademarks where a product or service meets a particular standard or has particular characteristics, such as quality or composition.[11] In the United Kingdom, innovators are able to ask the Digital Regulation Cooperation Forum’s (DCRF) AI and Digital Hub a specific query to understand how DCRF regulatory requirements may apply to a product, service or business model. The advice the Hub provides is not legally binding, nor a certification that the product, service or model is compliant with the law.[12]

**Recommendation 2**

HTI recommends the mandatory guardrails legislation include a mechanism for an AI developer or deployer to seek review, by an independent authority such as a regulator, of its own assessment regarding whether an AI system or model falls within the definition of ‘high-risk AI’.

## Application of AI laws to defence and national security organisations

HTI opposes broad exemptions for defence and national security organisations or in the context of law enforcement. Offering a class of organisations a general exemption from the obligation to comply with a law, rather than providing for case-by-case, justifiable exemptions and exceptions, would undermine government accountability. High-stakes decision-making contexts require strong accountability measures because they often present especially serious human rights risks and risks to safety. AI use cases in defence and national security include anything from highly intrusive biometric surveillance tools; profiling and social scoring; and AI weapons, including lethal autonomous weapons.

Broad exemptions would prevent sensible requirements to ensure that such uses of AI are safe; including basic steps to test these systems, conduct risk management processes, and mitigate potential harms, such as discrimination or physical harms to civilians. Such an approach would also conflict with the Australian Government's stated intention to position itself as an 'exemplar' with respect to AI use.

As a general principle, laws that are intended to protect the community – especially those that aim to uphold human rights – should apply to all legal persons. A person should be permitted to derogate from those protections only to the minimum extent necessary to pursue a lawful objective. Under international law, the fact that an organisation has, among its functions, the aim to protect Australia's defence or national security, should not, on its own, excuse the organisation from its duty to uphold human rights.

If a defence or national security body considers that the application of proposed mandatory guardrails legislation would impose a disproportionate fetter on its ability to defend Australia or its national security, this argument should be made openly by the relevant body. As a general principle of international law, where such an argument is successfully made, the legislature should not respond by granting the defence or national security body a blanket exemption from the relevant law – that would limit or restrict human rights more than is necessary to achieve the body's legitimate function.

Instead, the legislature generally should establish an independent adjudicative process to consider a detailed application from defence or national security bodies and decide whether the application should be granted. Such applications should be made on a case-by-case basis, in a similar way to the operation of a warrant scheme.

For example, if a national security body considers that certain transparency requirements would compromise a lawful national security operation, it could make the argument to an independent authority that it should not be required to comply with the relevant requirements. In this situation, there generally should be an alternative oversight mechanism, such as the Independent National Security Legislation Monitor, whose institutional arrangements ensure the protection of sensitive and classified information.

A similar approach to that outlined above is taken in a number of existing Australian legislative schemes, with Box 3 providing a particular example relating to the *Work Health and Safety Act 2011.*

**Box 3: Exemptions under the *Work Health and Safety Act 2011***

The *Work Health and Safety Act 2011*(Cth) (WHS Act) enables the Chief of the Australian Defence Force and other heads of national security organisations, with the approval of the Minister for Employment and Workplace Relations, to declare specific exemptions by written instrument. Declarations may state that specific provisions of the Act do not apply, or apply subject to modifications, in relation to specified activities, specified members of these organisations, or specified classes of members.[13]

For example, there is a defence exemption in relation to overseas operations, that removes WHS requirements to provide ‘immediate’ notification of death and injury, and to preserve WHS incident sites. This is due to the nature of military activities such as armed conflict situations where dangerous incidents are common, immediate reporting is logistically challenging, and where defence does not have effective control over the territory to preserve sites.[14]

Such declarations do not exempt defence and national security agencies from the application of the WHS Act as a whole. The WHS Act also includes a specific requirement for defence and national security agencies to ‘take into account the need to promote the objects of the Act to the greatest extent consistent with the maintenance of Australia’s defence and national security’.[15]

**Recommendation 3**

HTI recommends that the proposed mandatory guardrails legislation not contain a broad exemption for defence and national security organisations. Where a defence or national security organisation considers that this proposed law should be subject to an exception, there should be an independent adjudicative process to consider an application and decide whether the application should be granted, and on what terms.

## Harms to First Nations people

**Question 2:** Do you have any suggestions for how the principles could better capture harms to First Nations people, communities and Country?

HTI emphasises the importance of ensuring that the principles capture AI-based harms to First Nations peoples, and defers to First Nations individuals and organisations on this point. In particular, HTI notes the work of the First Nations Digital Inclusion Advisory Group.[16]

A key means of ensuring that First Nations peoples are considered, and their rights respected, is for the Government and organisations to engage in genuine consultation and co-design processes when drafting relevant law and policy, and developing or deploying AI systems that may affect First Nations people and communities. This submission provides further detail and a recommendation on this in Part B in response to Question 9.

## Unacceptable use cases and technology-specific laws

**Question 4:** Are there high-risk use cases that government should consider banning in its regulatory response (for example, where there is an unacceptable level of risk)? If so, how should we define these?

As discussed above, most instances of 'high-risk AI' should be sufficiently clear from the application of a definition that sets out the relevant factors. There is value also in giving the relevant Minister the power to designate certain AI systems or broader technology that pose a high or unacceptable risk.

This designation should be made where development or deployment of the technology in question would pose an unjustified restriction on human rights. For example, some forms of facial recognition technology (FRT) purport to be capable of assessing an individual's emotional state and other sensitive characteristics such as one's sexual orientation. HTI considers that such technology would pose an unjustifiable restriction on human rights if deployed to make decisions with a legal or similarly significant effect.

HTI has developed a model law for FRT, summarised in Box 4. This type of law could sit alongside an overarching AI law. The approach taken to classifying risk in the context of FRT could also be adapted and applied to other AI technologies that are deemed to be inherently high risk but have some acceptable uses.

**Box 4: Regulating the development & use of facial recognition technology**

FRT is regulated very lightly in Australia. This technology can carry a high risk of harm, because it can identify and extract a wealth of biometric and other sensitive personal information about an individual, often without their knowledge or consent. Some FRT systems have exhibited technical and operational problems – including issues with accuracy and bias.[17] FRT may also be used as a surveillance tool, or deployed for illegitimate or illegal purposes.[18]

In its Model Law, HTI proposed an approach to classifying risk in the specific context of FRT, recognising that while some use cases are unacceptable, others are acceptable with the correct safeguards in place.

**Recommendation 4**

HTI recommends that the proposed mandatory guardrails law enable the relevant Minister to assess whether certain AI systems or technology pose an unjustified restriction on human rights. In that scenario, the Minister should be able to designate by legislative instrument the relevant AI system or technology as high risk or prohibited.

## Approach to GPAI

**Question 5:** Are the proposed principles flexible enough to capture new and emerging forms of high-risk AI, such as general-purpose AI (GPAI)?

**Question 6:** Should mandatory guardrails apply to all GPAI models?

**Question 7**: What are suitable indicators for defining GPAI models as high-risk? For example, is it enough to define GPAI as high-risk against the principles, or should it be based on technical capability such as FLOPS (e.g. 10^25 or 10^26 threshold), advice from a scientific panel, government or other indicators?

HTI does not yet have a concluded view on the best regulatory response to GPAI. We look forward to continuing to engage with the Government on this issue, as well as with leading experts, like Gradient Institute, who have particular specialisation in this area.

# Part B: Guardrails ensuring testing, transparency and accountability of AI

**Question 8 (a):** Do the proposed mandatory guardrails appropriately mitigate the risks of AI used in high-risk settings?

**Question 10:** Do the proposed mandatory guardrails distribute responsibility across the AI supply chain and throughout the AI lifecycle appropriately for example other requirements assigned to developers and deployers appropriate?

In their current form, the proposed mandatory guardrails represent a concise and reasonably practical list of risk mitigation measures that can be undertaken by organisations to lessen the likelihood and extent of harms caused by high-risk AI systems. HTI sees value in this approach.

In particular, we endorse the focus on human-centred protections and accountability, including:

- publication of governance measures
- strengthening internal capability
- ensuring human oversight
- informing end users
- establishing redress processes for affected individuals.

A number of these safeguards could be improved through clarification of proposed requirements, as well as the addition of further details. HTI's response to questions 8(a) and 10 is addressed together for each guardrail and presented in the table below.

After this table, two additional mandatory guardrails are recommended.

**Table 2: HTI's input addressing questions 8(a) and 10 of the Proposals Paper**

| No. | Guardrail | HTI input addressing Q.8.a) and Q.10 of the Proposals Paper |
|---|---|---|
| 1 | *Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance* | While both developers and deployers should have clear governance processes to evaluate and mitigate any new or future harms that may eventuate due to their AI models or systems (for example, arising from changes in user demographics), *deployers* will need to have a clear focus on systems monitoring and ongoing harm prevention in the *specific deployment context of the AI system.*<br>Additionally, *both developers and deployers* should be required to outline a strategy for regulatory compliance, and document details of training provided to staff members. Specifically, there should be a requirement for all staff using high-risk AI systems to receive training on safe and responsible AI deployment, regardless of whether they are from a 'developing' or 'deploying' organisation. This is particularly important for deployers of high-risk AI systems that are human facing. |
| 2 | *Establish and implement a risk management process to identify and mitigate risks* | The guardrails should note that several versions of risk management processes might be required if the AI system could be reasonably deployed for a number of use cases and/or in varied environments.<br>It may also be appropriate for *deployers* to have access to risk management advice and processes established upstream by the developer, based on their initial classification of the AI system as high risk. |
| 3 | *Protect AI systems, and implement data governance measures to manage data quality and provenance* | Harms caused by poor data collection practices can arise at multiple points of the AI lifecycle, and manifest in different ways.<br>An illustrative example of this phenomenon was the Information Commissioner's 2021 determination that the Australian Federal Police (AFP) had breached privacy law in its use of Clearview AI's facial recognition tool. In its initial determination, the Information Commissioner focused on Clearview AI as the AI developer, finding that the company breached privacy law through its non-consensual scraping of sensitive biometric information from the web to train its FRT tool.[19] The Commissioner then considered the AFP effectively as a deployer of this FRT tool, and |

| No. | Guardrail | HTI input addressing Q.8.a) and Q.10 of the Proposals Paper |
|---|---|---|
| | | found that even though it was Clearview AI that had created the tool, the AFP bore responsibility as a deployer for ensuring that the deployment complied with federal privacy law.[20]<br>With this in mind, Guardrail 3 should make explicit the responsibility of an AI deployer within Australia to make reasonable efforts to satisfy itself that data collection and governance measures have been fulfilled by the relevant AI developer in accordance with Australian privacy and other relevant law.<br>AI developers should assist deployers in fulfilling this requirement by providing sufficient information on the nature, provenance and processes used to create training datasets that underpin their AI models and systems. While this information should ideally be public, it must be provided at least to deployers. |
| 4 | *Test AI models and systems to evaluate model performance and monitor the system once deployed* | To streamline testing and evaluation processes and improve market confidence (both among deployers and end-users of AI systems), developers should be primarily responsible for all initial performance testing and evaluation of an AI system for the general contexts it is designed to be deployed in.<br>Based on these assessments, performance assurance guarantees accompanied by approved general terms of use can be created to inform potential deployers of the system. Deployers should then be able to rely on these test guarantees if they deploy the AI system in ways that align with their initial design and terms of use. However, if a deployer significantly alters the model or diverts from the intended use case, the deployer will need to undertake its own performance testing processes.<br>More work is needed to determine how developers and deployers should bear their respective responsibilities for monitoring system performance on an ongoing basis. Just as manufacturers of products such as cars have an ongoing responsibility to monitor and act on safety and performance issues that arise after the relevant product has been sold, AI developers should continue to bear some monitoring responsibility post-deployment. However, the nature and extent of that responsibility would need to be determined by reference to the developer's limited capacity to see the ongoing operation of an AI system once different people have begun to deploy the system. |
| 5 | *Enable human control or intervention in an AI system* | This guardrail is one of the most important for ensuring the integrity and accountability of AI systems. It is well documented that if a human lacks the appropriate domain knowledge or technical skills, they will be less able or likely |

| No. | Guardrail | HTI input addressing Q.8.a) and Q.10 of the Proposals Paper |
|---|---|---|
| | *to achieve meaningful human oversight* | to remedy output errors of AI systems.[21] The Proposals Paper rightly emphasises that those responsible for the oversight of AI systems must be sufficiently qualified to interpret outputs and understand core capabilities and limitations of the model.<br>In national consultations undertaken by the Australian Human Rights Commission, stakeholders stressed the importance of humans in overseeing, monitoring and intervening in AI-informed decision making. With the right expertise, and oversight and governance arrangements, humans are well positioned to identify operating errors, exercise discretion in sensitive decision-making contexts, and assess the overall fairness and human rights compliance of an outcome.[22]<br>To achieve these outcomes – and in the pursuit of *meaningful* human oversight, as the guardrail suggests – deployers should either:<br>• have proof that staff have expertise in human-centred or responsible AI principles and be able to apply them to the AI systems they are overseeing, or<br>• provide adequate training to upskill their staff with this expertise.<br>This also complements the requirement in Guardrail 1 to develop accountability processes for ensuring internal capability. |
| 6 | *Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content* | This proposed guardrail promotes more transparent deployment of AI. Transparency can be a useful end in itself, but it is more commonly of value as a means of promoting greater accountability in a decision-making system or similar. Hence, HTI proposes that this guardrail be refined to ensure that end users are informed of when they are interacting with an AI system. This should be especially prioritised in circumstances where it is not readily apparent that AI is being used.<br>Where AI is being used to make a decision that has a legal or similarly significant effect on a person, this guardrail also should require: |

| No. | Guardrail | HTI input addressing Q.8.a) and Q.10 of the Proposals Paper |
|---|---|---|
| | | 1. AI developers and deployers to be capable of providing an explanation for the decision<br>2. AI deployers to inform affected people how they can seek human review of any resultant decision that has a legal or similarly significant effect, referring to the review processes covered by Guardrail 7. |
| 7 | *Establish processes for people impacted by AI systems to challenge use or outcomes* | In consultations undertaken by the Australian Human Rights Commission, there was broad agreement that people affected by AI-informed decision making should be entitled to have those decisions reviewed independently.[23] Guardrail 7 should make clear that relevant AI developers and deployers should provide *genuine opportunities for review and remedy*.<br>It is important to note that internal avenues for complaints also need to be supplemented by a strong regulatory ecosystem and potential law reform to improve redress options. This includes the ability for people to bring external complaints to regulators (such as consumer complaints via the ACCC), seek legal remedies through courts, and to enable responses to systemic issues that are flagged through complaints and review processes. |
| 8 | *Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks* | HTI endorses this proposed guardrail. |
| 9 | *Keep and maintain records to allow third parties to assess compliance with guardrails* | HTI endorses this proposed guardrail. |
| 10 | *Undertake conformity assessments to demonstrate* | HTI endorses this proposed guardrail and observes that significant additional work is needed to determine how conformity assessments should be undertaken and audited. Unlike other forms of conventional audits, there is still much uncertainty globally and no single accepted process for effective auditing of the safe development and |

| No. | Guardrail | HTI input addressing Q.8.a) and Q.10 of the Proposals Paper |
|---|---|---|
| | *and certify compliance with the guardrails* | deployment of AI. For example, international AI management standard ISO/IEC 42001:2023 was published less than a year ago. This standard has helped make clearer the criteria that companies should be assessing against, but precisely what it means to comply with this standard is still being determined. While the proposed Guardrail 10 would require organisations to prove adherence to the other guardrails, exactly what is required in order to demonstrate conformity for each of the guardrails will require detailed guidance.<br>Additionally, with no industry in Australia having developed sufficient maturity in AI audit functionality, the success of this guardrail will turn on the effectiveness of audit and regulatory oversight mechanisms. It would not be sufficient for compliance certification for the mandatory guardrails to be only a self-assessment process. A possible approach would be to require conformity assessments and the awarding of compliance certification to be undertaken by:<br>1. a third-party auditor – for the majority of high-risk systems<br>2. the relevant regulator – for very high-risk AI systems, perhaps within a particular subset of use cases.<br>In short, while proving compliance with the guardrails is a critical part of the AI risk mitigation process, the above issues will need to be addressed. The example of New York City Local Law 144 (passed in 2021) highlights the risks of mandating audits or compliance assessments relating to AI or automated decision making without sufficient clarity as to how they can be implemented effectively. Studies and reports since the commencement of that New York City law reveal widespread concerns about its effectiveness in achieving its aim of mandating impartial audits to curb bias in AI hiring algorithms, due to companies sidestepping requirements and failing to publish audits with poor results.[24] As a result, other jurisdictions looking at enacting similar auditing laws are now reconsidering their effectiveness. |

**Recommendation 5**

HTI recommends a number of refinements to the current draft of the mandatory guardrails (presented in Table 2), including further detail and clarification of the responsibilities between developers and deployers.

## Addition of two new mandatory guardrails

**Question 8 (b)** Are there any guardrails that we should add or remove?

### Have a plan for the safe decommissioning of AI systems

Risks exist across all stages of the AI lifecycle – from ideation, development, deployment, optimisation and decommissioning of an AI system. This final decommissioning or retirement stage is not commonly addressed during risk mitigation considerations, however, it raises particular issues that warrant safeguards. HTI recommends a mandatory guardrail be added to ensure developers and deployers have a plan in place for the safe and responsible decommissioning of AI systems.

An AI decommissioning plan should address issues such as:

- **Data handling and privacy –** notably, for the safe and appropriate deletion or sanitisation of information, including any data in back-ups, logs or external systems.
- **Knowledge preservation** – while data may need to be deleted, records and information about the system should be preserved for future reference and to meet any record-keeping or archives requirements.
- **System dependencies** – ensuring that the decommissioning process does not negatively affect other integrated systems which rely on the AI model or data
- **Transition or replacement plans –** for AI systems which underpin any critical infrastructure or services.[25]
- **Intellectual property and licensing** – managing the ownership or licensing of any AI systems, algorithms and datasets through appropriate termination or transferral.
- **Stakeholder communication** – all relevant parties should be informed with reasonable notice about the retirement of the AI system. Stakeholders may include internal staff such as management, data privacy teams, technicians and employees who rely on the system to do their work. External stakeholders requiring notice may include customers, clients and members of the public who engage with the system to access goods or services. Particular consideration will be required to notify and, where appropriate, compensate (or provide an alternate service to) customers who have paid for access to an AI-enabled system – for example, through a purchased health monitoring app on a smartphone or a medications subscription service that provides a personalised account with saved information.

### Considerations of the need for a ‘kill switch’

As part of a requirement for an AI decommissioning plan, there *may* be merit in provision for an emergency protocol for rapidly switching off an AI system. Sometimes known as a ‘kill switch’, the purpose of this safeguard is to ensure a last line of defence for models and/or systems that are operating dangerously or unpredictably and causing (or likely to cause) extreme harms at scale. This emergency mechanism is often conceptualised in the context of the ‘control problem’ of AI, which queries the limitations of effective human control over AI systems with advanced capabilities and their capacity to stay aligned with human-centred interests.[26] This theory includes the possibility that an AI system may, in fact, obstruct human efforts to curb its operation.

Following the release of DISR’s Proposals Paper, Minister Husic has referred to the need for human intervention in the safe and responsible use of AI in Australia, including consideration of the need for a ‘kill switch … if the AI that is being deployed is operating in a way that is not in line with what was expected’.[27]

While not using this terminology, a similar mechanism to a kill switch was presented in the Frontier AI Safety Commitments agreed to during the May 2024 AI Safety Summit in Seoul. Here, the emergency shutdown mechanism involves setting a threshold of unacceptable risk and requires a process to be in place to address that risk should the threshold be reached – including a commitment not to deploy the model or system if mitigations cannot be applied.[28] However, there is little clarity on what this threshold should be. Similarly, California’s SB-1047 Safe and Secure Innovation for Frontier Artificial Intelligence Models Bill (vetoed by Governor Gavin Newsom on 30 September 2024) contained provisions requiring organisations to be able to undertake a “full shutdown” of models and derivatives if they do not meet particular safety standards.[29]

On the other hand, HTI acknowledges that the idea of a kill switch is premised on AI models and systems being allowed to reach a point of such risk to humans that there is no other governance solution available to mitigate the relevant risk. This is inconsistent with the ‘whole of AI lifecycle’ risk management approach to the development and deployment of safe and responsible AI. In their work ‘The “Big red button” is too late,’ Arnold and Scheutz present an approach to identify and mitigate AI control problems prior to the point where a system has ‘gone rogue’ – including the AI system undertaking continued diagnostics and ‘opaque’ self-examination.[30] Pairing this kind of approach with rigorous testing and external oversight by humans from the start is likely the optimal way to manage risks before they reach a crisis point which can only be solved via a kill switch.

Ultimately, while HTI recommends the Government include an additional guardrail mandating a plan for the safe decommissioning of AI systems, we encourage the government to further deliberate the value and need of including a kill switch requirement as part of this. In addition, as noted in Recommendation 4, HTI recommends that the relevant Minister be given the power to determine that certain AI systems, models or technologies constitute a high or unacceptable risk, by reference to their likely restriction of human rights.

## Engage with stakeholders and evaluate their needs and circumstances

The impacts of AI are not dispersed or experienced equally by everyone. New technologies can, and do, lead to disproportionate harms, especially for people who already experience inequality or vulnerability, such as children, people with disability and Aboriginal and Torres Strait Islander Australians. Engaging with people to understand the potential impacts of AI on end users or affected individuals is essential to any human-centred approach to AI development and deployment.[31]

None of the ten proposed guardrails refers expressly to stakeholder engagement. This is the *only* significant difference between the proposed mandatory guardrails and the new Voluntary AI Safety Standard, which includes a standard to 'engage stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness'.[32] This Standard highlights that consultation with potentially impacted individuals and groups is a critical step in effectively identifying and mitigating risks of AI in order to maximise benefits for end users. More broadly, there is widespread acknowledgement of the importance of building trust and social licence through public involvement in the design stage of AI systems.[33]

Stakeholder engagement can also promote more accessible and inclusive design, development and deployment of AI. This is encapsulated by the 'nothing about us without us' rights-based approach popularised by modern disability rights movements and adopted in other deliberative processes, including AI development and deployment.[34] Stakeholder engagement can also serve commercial interests too. A 2023 study undertaken by the UK's Ada Lovelace Institute revealed that, within some commercial AI labs, public participation is considered to be positive both for society *and* for business.[35]

On the other hand, it is acknowledged that for the private sector — and especially for small and medium-sized enterprises (SMEs) — engaging in *genuine* stakeholder consultation is time consuming, resource intensive and challenging to do adequately and respectfully. There are legitimate, practical concerns regarding who should be consulted, and how to communicate with and recruit these individuals. Additionally, concerns regarding stakeholder exploitation and 'participant washing' are documented.[36]

Weighing these considerations, HTI recommends that consultation be a mandatory guardrail for development and deployment of high-risk AI by government. Given that government must act in the public interest, and its particular role in delivering critical services and entitlements to often at-risk or vulnerable groups, engagement with stakeholders should be mandatory. There are clear benefits of this for government with consultation being an important method for building community trust in, and uptake of, government policies and services. Governments should be well placed to fulfil this requirement, with stakeholder engagement a common feature of public service.[37]

In addition, while recognising that industry is not bound by the same obligations to deliver public good in the way that government is, there would be value in considering a requirement on the private sector also to undertake stakeholder consultation in respect of high-risk AI systems. There would be many ways of imposing such a requirement on the private sector, including through an additional guardrail, or by expanding one or more of the existing proposed guardrails. For example:

a. Guardrail 2: as part of establishing a risk management process, guidance could advise that consultation be undertaken when developers or deployers are considering 'any potential impacts on people, community groups and society before the high-risk AI system is in use.'[38]

b. Guardrail 4: in order to 'test the accuracy of an AI model for different social groups who may interact with an AI system to gauge the potential for discriminatory impacts',[39] guidance could recommend that developers and deployers undertake trials with community stakeholders who have consented to taking part.

Adding an explicit requirement for organisations to consult within the context of fulfilling these two guardrails would especially help to mitigate harms when a high-risk AI system is likely to be deployed on, or used by, someone from an at-risk group.

It should also be noted that a requirement to consult would not be unique to the AI context. Engagement with citizens, employees and customers is an established accountability mechanism in several areas, including workplace safety, employment law, environmental law and consumer rights.

**Box 5: Case study – a requirement to consult**

Under legislation, and standard clauses in all modern awards and enterprise agreements, employers have a duty to consult with employees when considering implementing significant workplace changes likely to impact their staff.[40] Employers must consult on issues like relocation, retraining and redundancies.[41]

In addition, the *Work Health and Safety Act 2011* (NSW) ('***WHS Act***') requires that:

- 'relevant information' about matters is shared with workers;
- workers are granted 'reasonable opportunity' to 'express their views' and 'contribute to the decision making process'; and
- that health and safety representatives are included in consultation processes.

The Fair Work Ombudsman provides best practice guidance for approaching meaningful consultation with employees.

There is an argument that minimum requirements for consultation give employers the flexibility to approach consultation in the way which best suits each workplace context and proposed change.[42] This less-prescriptive approach may well be transferable to the context of developer and deployer consultation with consumers and affected individuals in relation to high-risk AI uses.

**Recommendation 6**

In addition to the refinements to the Proposals Paper's current list of mandatory guardrails, HTI recommends that the Government consider the following additional mandatory guardrails:

a. a requirement to develop a plan for the safe decommissioning of high-risk AI models or systems.

b. a requirement for government to engage with stakeholders and evaluate their needs and circumstances in the development and deployment of high-risk AI models or systems

c. a requirement on the private sector to undertake stakeholder consultation in respect of high-risk AI systems, either via an additional mandatory guardrail or by expanding one or more of the existing proposed guardrails.

## Minimising AI harms to First Nations peoples, cultures and knowledge

**Question 9:** How can the guardrails incorporate First Nations knowledge and cultural protocols to ensure AI systems are culturally appropriate and preserve ICIP?

*HTI's advice to question 9 should be read in conjunction with our recommendation in response to question 8.b) to create a conditional requirement for stakeholder engagement in certain contexts of high-risk AI development and deployment.*

As AI technologies become increasingly integrated into decision making processes in Australia, including in particularly high-stakes contexts like health, justice and welfare, it is imperative that special consideration be given to the disproportionate impacts these systems may have on First Nations peoples, cultures and knowledge. In efforts to minimise these risks, consultation can help organisations better understand:

- how data on First Nations peoples, languages and cultures may need to be integrated into – or excluded from – the design of these models and systems (including via principles of Indigenous data sovereignty and governance)
- any racial, cultural or other biases demonstrated in the model
- possible negative or disproportionate impacts these systems may have on First Nations peoples.

In addition, the rise of generative AI systems poses unique challenges for First Nations Australians, particularly by way of Indigenous Cultural and Intellectual Property (ICIP). When large training datasets are created through poorly governed internet scraping practices, there is a risk that personal information related to Aboriginal and Torres Strait Islander people could be collected en masse and reconstituted in ways which lack cultural authority, permissions, sensitivity, accuracy, or attribution.[43]

By their very nature, generative AI systems do not possess the cultural wisdom or connection to country which is integral to the creation or sharing of First Nations art, stories and knowledge. The generation of synthesised AI outputs, like 'Indigenous' artworks, can therefore cause offence and distress, and breach a number of rights underpinned by the United Nations Declaration on the Rights of Indigenous Peoples (for example, Articles 8, 11 and 13). Again, actively including First Nations peoples in the planning stages of data collection and training could help reduce cultural exploitation, and move away from collecting data *about* these groups, rather than *for* and *with* them.[44] In the words of Maori ethicist, Karaitiana Taiuru, "Data is like our land

and natural resources. If Indigenous peoples don't have sovereignty of their own data, they will simply be re-colonized in this information society."[45]

Adopting HTI's recommended guardrail for stakeholder engagement would be a clear way to ensure First Nations peoples and cultures are taken into account in the context of high-risk AI systems which are likely to interface with or impact the legal rights of these Aboriginal or Torres Strait Islander Australians. Emphasis should be placed on genuine engagement through proper, culturally-appropriate consultation.

For public sector organisations, measures could go further, including ensuring alignment with the new Framework for Governance of Indigenous Data, which includes four guidelines:

1. Partner with Aboriginal and Torres Strait Islander people
2. Build data-related capabilities
3. Provide knowledge of data assets
4. Build an inclusive data system.[46]

Finally, there is an opportunity for the government to tie any advice it receives to Question 2 of this Proposals Paper to the practical measures sought here in Question 8. In other words, if the principles that determine whether an AI system is high risk adequately identity possible harms to First Nations people, communities and Country early in the design or pre-deployment stages, then these identified harms and their alleviation could form the basis of consultation in the risk mitigation process (that is, enacting the mandatory guardrails).

**Recommendation 7**

HTI recommends that the Government reduce the risk of AI systems harms to First Nations peoples, languages, cultures and knowledge by enacting a mandatory guardrail for consultation in particular high-risk contexts (per Recommendation 6).

## Reducing the regulatory burden on small-to-medium sized businesses

**Question 12:** Do you have suggestions for reducing the regulatory burden on small-to-medium sized businesses applying guardrails?

The recent release of the 2024 Australian Responsible AI Index report revealed that just 23% of surveyed organisations had implemented specific oversight and control measures to adequately govern AI systems.[47] This figure is likely to be smaller when looking at the AI governance activities of SMEs, given other research that shows the particular barriers SMEs face in implementing AI safeguards and governance processes. These barriers include availability of resources and skills; limited understanding of AI governance issues; lack of appropriate frameworks; and reluctance to engage with stakeholders on the design and use of new products and services.[48]

It is critical, therefore, that SMEs are supported to comply with the proposed guardrails. SMEs represent around 95% of Australian businesses,[49] and, currently, there is a general exemption for SMEs from compliance with the *Privacy Act 1988* (Cth). This leaves millions of Australians unprotected from many forms of privacy breach. While some sectors may be concerned about the compliance burden of extending mandatory guardrails to SMEs, the proposed legislation would apply only to high-risk AI and would likely not apply in many of the most common, day-to-day applications of AI by SMEs, such as for data entry, fraud detection and marketing automation.

**Box 6: Case study – SMEs**

A 2023 UK study[50] into practical solutions for the deployment of trustworthy AI systems by SMEs revealed that small businesses are looking for:

- **principles that are simple and flexible**, but more detailed than a checklist
- **practical guidance** on how to apply data and ethical AI principles
- **mechanisms, including training**, to support implementation
- **access to 'resource knowledge sharing'** for effective, ethical use of AI and machine learning
- **stakeholder involvement** in the development of data-driven technology strategies.

To promote effective implementation of the proposed mandatory guardrails, it will be necessary for government to provide clear, practical guidance to organisations on *how* to comply with the guardrails. Given the proposed guardrails closely mirror those of the Voluntary AI Safety Standard, advice and examples from this document could be adapted as the basis of guidance for the finalised mandatory guardrails.

Additionally, the Governance Institute of Australia, in collaboration with the National AI Centre, has released a White Paper providing leadership insights into the Voluntary AI Safety Standard in practice.[51] This document provides a number of 'expert tips' distilled from roundtable discussions with industry representatives, including for SMEs.

Importantly, SME resources, training and implementation advice will need to come from an authoritative body with appropriate funding to prioritise this. This organisation could be an existing regulator such as the ACCC, an appropriate peak body or NGO like the Council of Small Business Organisations Australia (COSBOA), or a newly established government AI Safety Commission (or similar) which can assist with capability uplift across different sectors.

**Recommendation 8**

HTI recommends that the Government provide clear, practical guidance to organisations on how to comply with the mandatory guardrails – including targeted resources and support for SMEs. This support should come from an appropriately resourced regulator or peak body.

# Part C: Regulatory options to mandate guardrails

**Question 13:** Which legislative option do you feel will best address the use of AI in high-risk settings? What opportunities should the government take into account in considering each approach?

**Question 15:** Which regulatory option/s will best ensure that guardrails for high-risk AI can adapt and respond to step-changes in technology?

## Overarching comment on the legislative options

The Proposals Paper states the aim of the mandatory guardrails is ‘to reduce the chance of harms occurring from the development and deployment of AI systems’, in order to ‘build trust and confidence in the use of such systems’.[52] There are various objectives for the legislative options set out in the Proposals Paper, such as setting ‘clear expectations’ from the Australian Government on safe and responsible AI use.[53] Option 2, for example, ‘would set a stronger signalling’ of the Government’s regulatory expectations and ‘represent best practices’ that agencies and regulators would be expected to build into their own domain specific laws.[54]

The Government’s legislative approach should incentivise safe and responsible development and deployment, empower regulators to enforce relevant legal obligations that relate to the development and deployment of AI, and provide accessible redress to individuals adversely affected by an AI system or model.

As the Government’s Interim Response and the Proposals Paper both acknowledge, the need to address the risk of harm from high-risk AI is immediate. Global technology companies are already taking advantage of weaker legal protections in certain jurisdictions, including Australia. Previously, the so-called ‘Brussels Effect’ enabled Australians to benefit from stronger protections as global companies uniformly altered their operations to comply with EU laws.[55] This is no longer the case. Meta recently confirmed to the Australian Parliament, for example, that it has enabled its European customers to opt out of Meta using their personal data to train their generative AI models but has not made this change in Australia, given that there is no legal requirement to do so.[56]

## Mandatory guardrails legislation should build on existing laws and enforcement mechanisms

As noted in the Proposals Paper, technology-neutral law already applies to high-risk AI. There is, however, a need to clarify how existing laws apply to the context of AI, as well as a need to achieve cross-sectoral clarity and consistency on a range of AI-related concepts.

Clarity is also needed on the attribution of liability across different stages of the AI lifecycle, and how current laws address liability. It is not clear, for example, how the Australian Consumer Law defences will apply to an AI product or system, given the interdependencies of AI components and the fact that a defect, or harm, may arise only after the point of supply.[57]

There are several domain-specific law reform or consultation processes underway, which are considering the implications of the rise of AI for existing law. Reforms, for example, have been agreed to, or are being considered, in the areas of privacy, copyright, consumer and administrative law.[58]

These reforms alone, however, will not be sufficient to address the harms associated with high-risk AI, and it is unlikely that a piecemeal reform approach would achieve a coherent regulatory response. There is also a risk of inconsistencies, and the potential for unnecessary complexity and confusing overlap. In the meantime, AI harms are likely to proliferate due to a lack of overarching safeguards, while uncertainty for business will continue.

Each of the legislative reform options outlined in the Proposals Paper would interact with existing laws in some way. HTI favours enactment of mandatory guardrails legislation that builds on our existing suite of laws and enforcement system. We consider that well-drafted legislation following the model of either Option 2 or Option 3 in the Proposal Paper could achieve this end. If Option 3 is adopted, HTI favours a model closer to Canada's draft AIDA Act, rather than the EU AI Act approach, because it is simpler and likely to be more straightforward to apply.

## Mandatory guardrails legislation should include a clear objective

Legislation for the proposed mandatory guardrails should articulate a clear and unambiguous objective. This objective should summarise the Government's aim in enacting this law – pointing to the outcomes that the law should help to achieve. The objects clause of the *Net Zero Economy Authority Act 2024* (Cth), for example, supports the Albanese Government's policy agenda to reduce emissions and make Australia a renewable energy superpower, while also ensuring support for regional areas and workers.[59]

HTI suggests a suitable objective for this proposed legislation would be to protect people from harm, and to support innovation for economic benefit and societal wellbeing.

Promoting innovation for societal 'wellbeing' aligns with the objectives of the Australian Government's national wellbeing framework, *Measuring what Matters*. This Framework tracks progress towards a more inclusive, equitable and fair society, and is intended to guide the priorities of both business and government decision making.[60] Linking the terminology in the regulatory objective with the *Measuring What Matters* framework will flesh out its substantive meaning and encourage consistency across the whole of government.

## Compliance and liability

### Obligation to take reasonable steps to comply with the mandatory guardrails

Laws specify, with varying degrees of precision, what people must do to comply with a particular legal requirement. Some laws require strict adherence to a legal requirement. Other laws offer some leeway for those who are covered by the law, such as by requiring a person to do what is considered 'reasonably necessary' to comply with the law.

A 'reasonable steps' requirement is common and generally well understood. The *Online Safety (Basic Online Safety Expectations) Determination 2022* (Cth), for example, requires that the provider of an online service 'take reasonable steps' to ensure the service can be used safely and to proactively minimise unlawful or harmful material or activity on the service.[61] Similarly, under the *Work Health and Safety (WHS) Act 2011* (Cth), duty-holders must do what is 'reasonably practicable' to fulfil the duty to ensuring health and safety.[62] This is an objective test, requiring a duty-holder to first consider what is possible in the circumstances for ensuring health and safety, and then take those steps unless it is reasonable in all the circumstances to do something less than that standard.[63]

HTI considers that it would be appropriate to require AI developers and deployers to take reasonable steps to comply with the mandatory guardrails.

### A rebuttable presumption regarding liability

A common concern regarding AI is that complex supply chains can make it difficult to determine who is liable in the event that a person suffers unjustified harm as a result of an AI system. The proposed mandatory guardrails legislation could mitigate this problem by providing for a 'default' position regarding liability.

This issue was addressed by the Australian Human Rights Commission's 2021 report on human rights and technology. The Commission recommended the creation, in law, of a rebuttable presumption, which would have the effect that 'where a corporation or other legal person is responsible for making a decision, that legal person is legally liable for the decision regardless of how it is made, including where the decision is automated or is made using artificial intelligence'.[64] In making this recommendation, the Commission noted this presumption should be no more than a general rule that could be displaced if there are strong legal reasons to do so.[65]

Adopting the language of the Proposals Paper, the effect of legislating a rebuttable presumption of this nature would be to set the default position that an AI deployer is liable for unjustified harm resulting from the operation of an AI system the deployer uses to make decisions that have a legal or similarly significant effect. Where the AI deployer can show that, in fact, another person was responsible for the relevant harm, such as an AI developer, the presumption would no longer apply (ie, the presumption will have been rebutted) and that other person will be liable. Crucially, however, it would be the responsibility of the AI deployer – and not a person suffering harm from the deployer's AI system – to reassign liability.

The use of a rebuttable presumption to assist in the apportionment of liability in the AI context would not be entirely novel. For instance, the proposed EU AI Liability Directive uses a rebuttable presumption to address the challenges of apportioning liability for high-risk AI under the EU AI Act. This mechanism is intended to protect the right to an effective remedy under the EU Charter of Fundamental Rights, incentivise corporate behaviour to prevent harm, and protect innovation.[66] If passed, this Directive would introduce a rebuttable presumption that there is a causal link between the defendant's fault, and the output of an AI system or failure to produce an output. National courts should apply the presumption where three conditions are met: (a) the claimant has demonstrated that the defendant is at fault; (b) it is reasonably likely the fault influenced the AI system's output or failure; and (c) there is sufficient proof the output gave rise to damage.[67]

## Enforcement of the mandatory guardrails should be addressed in legislation

In order to make the guardrails mandatory, the proposed legislation will need to provide for appropriate enforcement. Option 3 contemplates having enforceable provisions, which would operate alongside existing avenues of redress in domain-specific laws. Enforcement of the guardrails under Option 1 would depend on reform of existing law; under Option 2, enforcement would depend on how the framework legislation is activated.

There are several possible enforcement mechanisms that could be adopted to address non-compliance with the mandatory guardrails. These mechanisms are not mutually exclusive. Parliament could legislate for one, or a combination of enforcement mechanisms. As explained below, HTI favours a combination of regulator-led enforcement and 'piggy-back' provision, as opposed to a new, direct cause of action.

### Regulator-led enforcement

This mechanism of enforcement would rely primarily on action by an oversight body. The regulator could. of its own motion and/or in response to a complaint from the public, investigate possible non-compliance with the guardrails. In the event that the regulator is satisfied that there has been a breach of the guardrails, the regulator could have a power to order remedial action and/or civil penalties.

This regulator could be any of the following:

- a single government regulator that is given sole responsibility to oversee compliance with the mandatory guardrails across the entire Australian economy. This could be a new regulatory body, or an existing regulator such as the Office of the Australian Information Commissioner
- a group of existing government regulators, which is tasked with overseeing compliance with the mandatory guardrails by organisations in the respective sectors for which those regulators are responsible[68]
- a private sector body, such as an industry ombudsman, which is tasked with overseeing compliance with the mandatory guardrails by organisations in the respective sectors for which those regulators are responsible.

## A new cause of action

The most stringent form of enforcement would be to create a new cause of action that would enable a person with standing (ie, a person who has a special interest) to sue an AI developer or deployer based solely on the developer's or deployer's non-compliance with the mandatory guardrails. This would likely involve the person taking this matter to a court or tribunal, which would be able to provide a remedy (such as damages) to address harm that is proven from the non-compliance.

Without expressing a concluded view on this possible enforcement mechanism, HTI has some reservations about the suitability of this option. HTI's concern with this enforcement mechanism stems from the fact that the mandatory guardrails outlined in the Proposals Paper essentially set out process-based requirements that are designed to reduce the likelihood of high-risk AI causing harm. Compliance with the guardrails does not guarantee protection from such harm, but neither does non-compliance make it certain that harm will arise.

A direct cause of action, founded on breach of one or more guardrails, could create a situation where a person (the plaintiff) is able to sue an organisation for failing to comply with the mandatory guardrails in circumstances where the plaintiff has not suffered any compensable harm.

## A 'piggy-back' provision

Unlike a new, independent cause of action, a 'piggy-back' provision would allow a person with legal standing, who already has a cause of action against an AI developer or deployer, to support their argument in respect of that existing cause of action by reference to the developer or deployer's non-compliance with the guardrails.

Sometimes referred to as a 'piggy-back' provision or clause, this approach means a court would consider whether one or more of the guardrails have been complied with where the operation of an AI model or system is relevant to another cause of action or pursuit of a judicial remedy.

Both the *Charter of Human Rights and Responsibilities 2006* (Vic) and the *Human Rights Act 2019* (Qld) contain examples of 'piggy-back' clauses. Section 39(1) of the Victorian Charter, for example, states:

> If, otherwise than because of this *Charter*, a person may seek any relief or remedy in respect of an act or decision of a public authority on the ground that the act or decision was unlawful, that person may seek that relief or remedy on a ground of unlawfulness arising because of this *Charter*.[69]

The 'piggy-back' clause in the Victorian Charter was a drafting solution to ensure individuals can access relief for human rights breaches, while complying with the Victorian Government's stated intention not to create a new individual cause of action based on a human rights breach.[70] It has been described in case law as 'an enabling provision', giving courts and tribunals the additional jurisdiction needed to give effect to, and vindicate, Charter rights:

> In cases where the lawfulness of an act or decision of a public authority may be challenged on an independent ground of unlawfulness, it allows a person to seek the available relief or remedy on a ground of unlawfulness because of the Charter.[71]

HTI considers that a carefully drafted piggy-back provision would be an appropriate, balanced enforcement mechanism for a person who suffers harm as a result of an organisation's development or deployment of a high-risk AI system, to demonstrate that the organisation failed to take appropriate steps to mitigate the risk of harm. A hypothetical example of how a piggy-back clause might apply to an AI-related harm is set out in Box 7 below.

**Box 7: Hypothetical example of the piggy-back provision**

Imagine that a bank contracts with a small business, Company X, to extend it a credit facility subject to the bank's assessment of Company X's credit worthiness. The bank proposed to make that assessment based on the outcome of an AI-based credit scoring tool. The AI tool has not been subjected to rigorous testing in accordance with proposed Guardrail 4. While the bank is not aware of this prior to using the AI tool, the bank's tool happens to be highly accurate when used to assess credit worthiness for large companies with a turnover greater than $10 million, but it is prone to high rates of errors for SMEs (and specifically higher rates of error than other credit scoring tools to which the bank has access).

On the basis of the AI tool, the bank assessed Company X as ineligible for the extended credit facility. This resulted in Company X suffering significant financial loss. Company X sued the bank, arguing the bank failed in a legal duty to 'exercise the care and skill of a prudent banker in selecting and applying our credit assessment methods', after it was discovered the AI tool was prone to high rates of errors when used in respect of small businesses like Company X.

In this scenario, if there was a piggy-back provision that enabled non-compliance with the mandatory guardrails to be adduced as evidence to support an existing cause of action, Company X would be able to argue that the bank's failure to conduct testing of the AI tool in respect of SMEs, before using the tool on Company X, supported its argument that the bank had breached its duty to 'exercise the care and skill of a prudent banker'.

## Other enforcement options

While the three mechanisms described above represent the most obvious options for enforcement, there are other possible options as well. For example, Parliament could choose to make non-compliance with the mandatory guardrails a criminal offence. This would mean that the officers of an AI developer or deployer could be found criminally liable for breaching one or more of the mandatory guardrails. However, other than in very serious cases of non-compliance with a regulator's or court's order, it seems that a criminal enforcement provision would be more punitive on AI developers and deployers than would be necessary to promote compliance with the guardrails.

**Recommendation 9**

If the Government introduces legislation containing a list of mandatory guardrails for developers and deployers of high-risk AI, the legislation should:

a. include a definition of 'high-risk AI', in the manner set out in Recommendation 1

b. set out a clear regulatory objective

c. require AI developers and deployers to take reasonable steps to comply with the mandatory guardrails

d. contain a rebuttable presumption that where a person is responsible for making a decision using AI, that person is legally liable for the impact of that decision

e. provide for enforcement through appropriate mechanisms such as oversight by a regulator, and a 'piggy-back' provision that would support people with an existing cause of action based on s 39(1) of the *Charter of Human Rights and Responsibilities 2006* (Vic).

## Incentivising compliance

While the realistic prospect of enforcement is important in promoting compliance with the proposed mandatory guardrails, the Government should also consider more positive incentives to encourage AI developers and deployers to see value in complying with the guardrails. While more detailed work would be needed to evaluate their suitability, some possible incentives include:

- a regulator or other authoritative body offering practical, sector-specific advice or guidance on steps needed to comply with the various guardrails
- government procurement rules being amended to prioritise companies that demonstrate compliance with the mandatory guardrails
- reducing the potential legal liability of an organisation that has demonstrated good-faith compliance with the mandatory safeguards.

It appears likely that some larger companies, especially AI developers, will choose to assist compliance with the mandatory guardrails by AI deployers to whom they provide AI systems. Typically, in this sort of scenario, an AI developer would offer terms of use for an AI system wherein the AI developer warrants compliance with the mandatory guardrails if an AI deployer uses the AI system in a certain way. The Government should consider options for supporting more efficient and effective compliance with the mandatory guardrails, especially by SMEs that are content to operate an AI system entirely within the relevant terms of use.[72]

**Recommendation 10**

HTI recommends the Government consider additional measures to support compliance with the proposed mandatory guardrails legislation, including:

a. guidance from a regulator or other authoritative body

b. amending procurement rules to prioritise companies demonstrating compliance with the mandatory guardrails

c. reducing, to an appropriate extent, the relevant legal liability of an organisation that has demonstrated compliance with the mandatory safeguards.

[1] Department of Industry, Science and Resources (Commonwealth), *Safe and responsible AI in Australia consultation* (Interim Response, 17 January 2024) 18 <https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf>.

[2] The principles include: adopting a risk-based framework; balanced and proportionate, avoiding unnecessary burdens for businesses, community and regulators; adopting a collaborative and transparent approach, including engaging with experts and ensuring public involvement; being a trusted international partner, consistent with Australia's commitments in the Bletchley Declaration; placing people and communities at the centre of its regulatory approach. See Department of Industry, Science and Resources (Commonwealth), *Safe and responsible AI in Australia consultation* (Interim Response, 17 January 2024) 18–9 <https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf>.

[3] OECD, *Common guideposts to promote interoperability in AI risk management* (OECD Artificial Intelligence Papers No. 5, November 2023) <https://www.oecd.org/en/publications/common-guideposts-to-promote-interoperability-in-ai-risk-management_ba602d18-en.html>; High-Level Advisory Body on Artificial Intelligence, United Nations, *Interim Report: Governing AI for Humanity* (Report, December 2023) <https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf>.

[4] *Human Rights (Parliamentary Scrutiny) Act 2011* (Cth) s 3(1).

[5] Adapted from: Parliamentary Joint Committee on Human Rights (Commonwealth), *Guidance Note 1: Drafting statements of compatibility* (Guidance Note, December 2014) <https://www.aph.gov.au/~/media/Committees/Senate/committee/humanrights_ctte/guidance_notes/guidance_note_1/guidance_note_1%20(4).pdf?la=en>; UN Commission on Human Rights, Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights, UN Doc E/CN.4/1985/4 (28 September 1984).

[6] 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 [2024] OJ L 2024 1689, art 6 <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.

[7] Art 3, *Artificial Intelligence Act*, Regulation (EU) 2024/1689.

[8] Australian Government, *Government response to the Privacy Act Review Report* (Report, September 2023) 11 <https://www.ag.gov.au/sites/default/files/2023-09/government-response-privacy-act-review-report.PDF>.

[9] Article 29, Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (February 2018) 21.

[10] Jarni Blakkarly, ‘Is Airbnb using an algorithm to ban users from the platform?’ *Choice* (Web Page) 21 March 2022 <https://www.choice.com.au/consumers-and-data/data-collection-and-use/how-your-data-is-used/articles/airbnb-banning-users>.

[11] Australian Competition and Consumer Commission ‘Certification trade marks’ (Web Page, accessed 3 October 2024) <https://www.accc.gov.au/business/competition-and-exemptions/exemptions-from-competition-law/certification-trade-marks>.

[12] Digital Regulation Competition Forum ‘AI and Digital Hub (Web Page, accessed 3 October 2024) <https://www.drcf.org.uk/ai-and-digital-hub/>.

[13] *Work Health and Safety Act 2011* (Cth) ss 12C–12E.

[14] Work Health and Safety Act 2011 (application to Defence activities and Defence members) Declaration *2012* (Cth) <https://www.legislation.gov.au/F2012L02503/latest/text>.

[15] *Work Health and Safety Act 2011* (Cth) ss 12C(4)–(6), 12D(4)–(5).

[16] ‘Submissions from the Advisory Group’*, First Nations Digital Inclusion Advisory Group* (Web Page) <https://www.digitalinclusion.gov.au/submissions>.

[17] Samuel Dooley et al., ‘Comparing Human and Machine Bias in Face Recognition’ (Research Paper, *arXiv Computer Vision and Pattern* Recognition, 15 October 2021) <http://arxiv.org/abs/2110.08396>.

[18] Nicholas Davis, Lauren Perry and Edward Santow, *Facial recognition technology: Towards a model law* (Report, September 2022) <https://www.uts.edu.au/sites/default/files/2022-09/Facial%20recognition%20model%20law%20report.pdf>

[19] Office of the Australian Information Commissioner, ‘Clearview AI breached Australians’ privacy (Media Release, 3 November 2021) <https://www.oaic.gov.au/news/media-centre/clearview-ai-breached-australians-privacy>.

[20] Office of the Australian Information Commissioner, ‘AFP ordered to strengthen privacy governance’ (Media Release, 16 December 2021) <https://www.oaic.gov.au/news/media-centre/afp-ordered-to-strengthen-privacy-governance>.

[21] For example, Jake Goldenfein ‘Algorithmic Transparency and Decision-making Accountability’, in Office of the Victorian Information Commissioner (ed), *Closer to the Machine: Technical, Social and Legal Aspects of AI* (August 2019) 56.

[22] Australian Human Rights Commission, Human Rights and Technology Final Report (Report 2021) 69, 71 <https://humanrights.gov.au/sites/default/files/document/publication/ahrc_rightstech_2021_final_report_10.pdf>

[23] Australian Human Rights Commission, Human Rights and Technology Final Report (Report 2021) 69 <https://humanrights.gov.au/sites/default/files/document/publication/ahrc_rightstech_2021_final_report_10.pdf>.

[24] Brandon Vigliarolo, ‘Law designed to stop AI bias in hiring decisions is so ineffective it's slowing similar initiatives’, *The Register* (online, 23 January 2024) <https://www.theregister.com/2024/01/23/nyc_ai_hiring_law_ineffective>.

[25] Information Commissioner’s Office, *Guidance on AI and data protection* (Regulatory Guidance, 13 March 2023) 133 <https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection-2-0.pdf>.

[26] Nik Samoylov, The “Control Problem” of advanced artificial intelligence (AI), (Fact Sheet, 4 June 2023) <https://www.campaignforaisafety.org/content/files/2023/06/The--Control-Problem--of-advanced-artificial-intelligence---factsheet.pdf>.

[27] Kat Wong and Jennifer Dudley-Nicholson, ‘‘Kill switch’ considered to tackle high-risk AI’, *Australian Associated Press* (online, 5 September 2024) <https://www.aap.com.au/news/kill-switch-considered-to-tackle-high-risk-ai/>.

[28] Department for Science, Innovation, & Technology (UK), *Frontier AI Safety Commitments, AI Seoul Summit 2024* (Policy Paper, 21 May 2024)

<https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024>.

[29] C.A. Sen. SB-1047. Reg. Sess. 2023-2024 (2024) <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047>.

[30] Thomas Arnold and Matthias Scheutz, 'The "big red button" is too late: an alternative model for the ethical evaluation of AI systems' (2018) 20 *Ethics and Information Technology* 59 <https://doi.org/10.1007/s10676-018-9447-7>.

[31] Human Technology Institute, *Putting people at the centre of AI – impacted communities and missing voices* (Snapshot, January 2024) <https://www.uts.edu.au/sites/default/files/2024-01/AI%20Governance%20Snapshot%20%232%20Putting%20people%20at%20the%20centre%20of%20AI%20%E2%80%93%20impacted%20communities%20and%20missing%20voices.pdf>.

[32] Department of Industry, Science and Resources (Commonwealth), *Voluntary AI Safety Standard* (August 2024) 42–4 <https://www.industry.gov.au/sites/default/files/2024-09/voluntary-ai-safety-standard.pdf>.

[33] Keeley Crockett, Edwin Colyer, Luciano Gerber and Annabel Latham, 'Building Trustworthy AI Solutions: A Case for Practical Solutions for Small Businesses' (2023) 4(4) *IEEE Transactions on Artificial Intelligence* 778 <https://ieeexplore.ieee.org/document/9658213>; Timothy Murphy, Swati Garg, Brenna Sniderman and Natasha Buckley, 'Ethical technology use in the fourth industrial revolution', *Deloitte Insights* (Article, 15 July 2019) <https://www2.deloitte.com/us/en/insights/focus/industry-4-0/ethical-technology-use-fourth-industrial-revolution.html>.

[34] ARC Centre of Excellence for Automated Decision-Making and Society, Submission No 437 to Department of Industry, Science and Resources, *Safe and Responsible AI Discussion Paper* (4 August 2023) 2 <https://apo.org.au/sites/default/files/resource-files/2023-08/apo-nid323896.pdf>.

[35] Lara Groves, Aidan Peppin, Andrew Strait and Jenny Brennan, 'Going public: the role of public participation approaches in commercial AI labs' (Conference Paper, ACM Conference on Fairness, Accountability, and Transparency, 16 June 2023) <https://arxiv.org/abs/2306.09871>.

[36] Lara Groves, Aidan Peppin, Andrew Strait and Jenny Brennan, 'Going public: the role of public participation approaches in commercial AI labs' (Conference Paper, ACM Conference on Fairness, Accountability, and Transparency, 16 June 2023) 8 <https://arxiv.org/abs/2306.09871>.

[37] For example, 'Getting stakeholder engagement right', *Australian Public Service Commission* (Web Page, 27 February 2024) <https://www.apsc.gov.au/initiatives-and-programs/aps-mobility-framework/taskforce-toolkit/stakeholder-engagement/getting-stakeholder-engagement-right>; Department of the Prime Minister and Cabinet, *Best Practice Consultation* (Guidance Note, July 2023) <https://oia.pmc.gov.au/sites/default/files/2023-08/best-practice-consultation.pdf>.

[38] Department of Industry, Science and Resources (Commonwealth), *Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings* (Proposals Paper, September 2024) 36 <https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf>.

[39] Department of Industry, Science and Resources (Commonwealth), *Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings* (Proposals Paper, September 2024) 38 <https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf>.

[40] Fair Work Ombudsman, 'Consultation and cooperation in the workplace' (Report, 2022) <https://www.fairwork.gov.au/tools-and-resources/best-practice-guides/consultation-and-cooperation-in-the-workplace#using-best-practice-to-support-consultation-and-cooperation-in-the-workplace>.

[41] Giuseppe Carabetta and Paul Lorraine, 'Meeting the obligation to consult in employment law', *LSJ Online* (online, 8 May 2024) <https://lsj.com.au/articles/meeting-the-obligation-to-consult-in-employment-law/>.

[42] Giuseppe Carabetta and Paul Lorraine, 'Legal Parameters of the Employer's Duty to Consult' (2023) 42(2) *University of Queensland Law Journal* 223 <https://www8.austlii.edu.au/au/journals/UQLawJl/2023/9.pdf>.

[43] Emma Fitch, Clare McKenzie, Terri Janke and Adam Shul, 'The new frontier: Artificial Intelligence, copyright and Indigenous Culture', *Terri Janke and Company* (Blog Post, 30 November 2023) <https://www.terrijanke.com.au/post/the-new-frontier-artificial-intelligence-copyright-and-indigenous-culture>.

[44] Bronwyn Carlson and Peita Richards, 'Indigenous knowledges informing 'machine learning' could prevent stolen art and other culturally unsafe AI practices', *The Conversation* (online, September 8, 2023) <https://theconversation.com/indigenous-knowledges-informing-machine-learning-could-prevent-stolen-art-and-other-culturally-unsafe-ai-practices-210625>.

[45] Rina Chandran, 'Indigenous groups fear culture distortion as AI learns their languages', *The Japan Times* (online, 10 April 2023) <https://www.japantimes.co.jp/news/2023/04/10/world/indigenous-language-ai-colonization-worries/>.

[46] Australian Government, *Framework for Governance of Indigenous Data* (Report, May 2024) <https://www.niaa.gov.au/sites/default/files/documents/2024-05/framework-governance-indigenous-data.pdf>.

[47] Fifth Quadrant, *The Australian Responsible AI Index 2024* (Report, September 2024) 21 <https://www.fifthquadrant.com.au/responsible-ai-index-2024>.

[48] Keeley Crockett, Edwin Colyer, Luciano Gerber and Annabel Latham, 'Building Trustworthy AI Solutions: A Case for Practical Solutions for Small Businesses' (2023) 4(4) *IEEE Transactions on Artificial Intelligence* 778 <https://ieeexplore.ieee.org/document/9658213>.

[49] Australian Small Business and Family Enterprise Ombudsman, *Number of small businesses in Australia* (Report, August 2023) 2 <https://www.asbfeo.gov.au/sites/default/files/2023-10/Number%20of%20small%20businesses%20in%20Australia_Aug%202023_0.pdf>.

[50] Keeley Crockett, Edwin Colyer, Luciano Gerber and Annabel Latham, 'Building Trustworthy AI Solutions: A Case for Practical Solutions for Small Businesses' (2023) 4(4) *IEEE Transactions on Artificial Intelligence* 778 <https://ieeexplore.ieee.org/document/9658213>.

[51] Governance Institute of Australia and National Artificial Intelligence Centre, *White Paper on AI Governance: Leadership insights and the Voluntary AI Safety Standard in practice* (Report, 11 September 2024) <https://www.governanceinstitute.com.au/thought-leadership/ai-ethics-and-governance-white-paper-launch/>.

[52] Department of Industry, Science and Resources (Commonwealth), *Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings* (Proposals Paper, September 2024) 30 <https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf>.

[53] Department of Industry, Science and Resources (Commonwealth), *Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings* (Proposals Paper, September 2024) 2, 30 <https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf>.

[54] Department of Industry, Science and Resources (Commonwealth), *Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings* (Proposals Paper, September 2024) 49 <https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf>.

[55] Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (Oxford University Press, 2020).
[56] Jake Evans 'Facebook admits to scraping every Australian adult user's public photos and posts to train AI, with no opt-out option' *ABC News* (online, 11 September, 2024) <https://www.abc.net.au/news/2024-09-11/facebook-scraping-photos-data-no-opt-out/104336170>.
[57] Commonwealth of Australia, *Consumer product safety: A guide for businesses and legal practitioners* (Guide, November 2022) <https://consumer.gov.au/sites/consumer/files/inline-files/acl-guidance-consumer-product-safety.pdf>. For detailed examination of the interaction of product liability and AI see, for example, Miriam Buiten, Alexandre de Streel and Martin Peitz, 'The law and economics of AI liability' (2023) 48 *Computer Law & Security Review* 1 <https://www.sciencedirect.com/science/article/pii/S0267364923000055>.
[58] Attorney-General's Department, *Privacy Act Review Report* (Report, 16 February 2023) <https://www.ag.gov.au/rights-and-protections/publications/privacy-act-review-report>; 'Copyright and Artificial Intelligence Group (CAIRG)', *Attorney-General's Department* (Web Page) <https://www.ag.gov.au/rights-and-protections/copyright/copyright-and-artificial-intelligence-reference-group-cairg>; *Royal Commission into Robodebt* (Final Report, 7 July 2023) <https://robodebt.royalcommission.gov.au/>.
[59] *Net Zero Economy Authority Act 2024* (Cth) s 3.
[60] 'Measuring What Matters', The Treasury (Web Page) <https://treasury.gov.au/policy-topics/measuring-what-matters>.
[61] *Online Safety (Basic Online Safety Expectations) Determination 2022* (Cth) s 9.
[62] *Work, Health and Safety Act 2011* (Cth) s 18.
[63] Safe Work Australia, *Model Work Health and Safety Act, the meaning of 'reasonably practicable'* (Interpretive Guideline, 28 October 2011) <https://www.safeworkaustralia.gov.au/doc/interpretive-guideline-model-work-health-and-safety-act-meaning-reasonably-practicable>.
[64] Australian Human Rights Commission, Human Rights and Technology Final Report (Report, 2021) 78 <https://humanrights.gov.au/sites/default/files/document/publication/ahrc_rightstech_2021_final_report_10.pdf>.
[65] Australian Human Rights Commission, Human Rights and Technology Final Report (Report, 2021) 80 <https://humanrights.gov.au/sites/default/files/document/publication/ahrc_rightstech_2021_final_report_10.pdf>.
[66] Explanatory Memorandum, 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) COM/2022/496 final, 9 <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A52022PC0496>.
[67] 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) COM/2022/496 final, art 4 <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0496>.
[68] For an example of an existing forum of regulators, and consideration of co-ordination and potential mandate overlap, see 'DP-REG Terms of Reference', *Digital Platforms Regulators Forum* (Web Page, 25 July 2024) <https://dp-reg.gov.au/dp-reg-terms-reference>.
[69] Cf. *Human Rights Act 2019* (Qld) s 59.
[70] Department of Justice (Vic), *Human Rights in Victoria: Statement of Intent* (Statement, 2005). See, also, George Williams, 'The Victorian *Charter of Human Rights and Responsibilities:* Origins and Scope' (2006) 30(3) *Melbourne University Law Review*, 880, 895-7 <https://www8.austlii.edu.au/cgi-bin/viewdoc/au/journals/MelbULawRw/2006/27.html>. It is noted that the drafting solution has been criticised for being ambiguous, and the limitations of not having an independent cause of action in human rights legislation: see, for example, Jeremy

Gans, 'The *Charter's* irremediable remedies provision' (2009) 33 *Melbourne University Law Review* 105 <https://law.unimelb.edu.au/__data/assets/pdf_file/0004/1705315/33_1_4.pdf>; Parliamentary Committee on Human Rights, *Inquiry into Australia's Human Rights Framework* (Report, May 2024) 7, 98 <https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Human_Rights/HumanRightsFramework/Report>.

[71] *Louise Goode v Common Equity Housing Ltd* [2014] VSC 585 [39] (Bell J).

[72] For a products liability and consumer products safety-type approach for high-risk AI, see Center for Humane Technology, *Framework for Incentivising Responsible Artificial Intelligence Development and Use* (Report, 12 September 2024) <https://www.humanetech.com/insights/framework-for-incentivizing-responsible-artificial-intelligence>.