Social Policy Group's Submission to the Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings

Table of Contents

The Social Policy Group (SPG) is a national, non-government, not-for-profit body with specialist expertise in social policy and program design with a focus on population diversity, social and community cohesion, gender equality, community participation and inclusion, systems' responsiveness, and community outreach and engagement. The Social Policy Group is a peak body representing settlement services, health as well as auspicing the Harmony Alliance representing migrant and refugee women. The Social Policy Group has significant expertise in AI systems and governance and has prepared this submission to draw attention to the inadequacy of existing and proposed approaches and the consequences for marginalised and vulnerable people and communities in Australia. The purpose of this submission is not to address the proposed questions but address the overall gaps and risks identified with Australia's proposed approach.

Addressing Gaps in a Framework for an Evolving Technological Landscape

The development of mandatory guardrails for AI in high-risk settings is a necessary recognition of the importance of regulating emerging technologies. However, AI is not a static tool, and its integration into Australian industries is marked by its evolving nature, global complexity, and increasing capacity for adaptation. It should be acknowledged that these factors create significant regulatory challenges. While a first step, the framework in its current form does not sufficiently account for the future trajectory of AI systems. By focusing on immediate technical and organisational concerns, the current approach risks falling short of its intended goals. A more nuanced, adaptive framework is required to prevent emergent gaps from undermining the integrity of AI governance.

AI, particularly when powered by neural networks and deep learning models, is inherently dynamic. These systems continuously learn from new data, adapt to changing environments, and evolve beyond their initial design. The challenge lies not in their capabilities but in the regulatory framework's capacity to account for their evolution. Categorising risk at the point of deployment assumes that these systems remain within predictable operational boundaries. This is not the case. The current proposal relies heavily on conformity assessments and internal organisational processes that may not adequately capture the complexities and risks that evolve over time. Moreover, general-purpose AI (GPAI) systems, which can be applied across a broad spectrum of use cases, further complicate this approach.

Neural Networks and Deep Learning: Evolving Risk and the Threat of Manipulation

Neural networks and deep learning models function differently from traditional software systems. These architectures are designed to absorb new data, refine their internal algorithms, and adapt their outputs based on ongoing interactions. This continuous learning process, while critical to their effectiveness, introduces an unpredictable risk trajectory. AI systems designed for one use case can evolve to handle more complex or sensitive tasks, even without human intervention, which raises significant challenges for static risk categorisation.

The risk lies not only in the system's capability to learn but also in its capacity to manipulate. Deep learning models, particularly those deployed in high-risk environments like financial services or content curation, are increasingly optimised for engagement and efficiency. In practice, this can result in AI systems subtly shaping user behaviour, decisions, and preferences. This manipulation, whether intentional or algorithmic, often occurs without the user's awareness, creating ethical and societal risks that are not adequately addressed within the current regulatory framework.

For example, AI systems in social media environments use engagement optimisation algorithms that present users with content designed to reinforce their interests and maximise time spent on the platform. While this may serve an organisation's objectives, the societal consequences—including the polarisation of discourse, the reinforcement of biases, and the spread of misinformation—are significant. The current guardrails focus primarily on technical risks to individuals but do not extend to the broader societal implications of these AI systems. Manipulation at scale can reshape public opinion and behaviour, yet the framework is silent on these risks.

Who Determines Risk? The Complexity of Context and Use Case

The proposed guardrails assign the responsibility for determining risk primarily to developers and deployers. This determination is contextual, depending heavily on the system's use case and the environment in which it is deployed. In high-risk settings, the expectation is that organisations will implement risk management processes before these systems are used, informed by a combination of AI impact assessments and principles set out in the proposal. The focus on organisational responsibility seems appropriate at first glance, but it also introduces critical gaps when it comes to general-purpose AI (GPAI).

GPAI systems, due to their adaptive nature, pose a challenge because they are not constrained by a single use case. These models can evolve and be applied in multiple contexts beyond the developer's or deployer's initial intentions. This means that determining risk based on an initial assessment is not enough; it requires continuous monitoring and reassessment as the system's capabilities grow. However, the framework's reliance on conformity assessments and initial risk sign-offs falls short in addressing these ongoing changes.

Furthermore, developers and deployers are given significant leeway to determine their own risk thresholds based on their organisational needs, which can lead to inconsistent application of risk management principles across sectors. Each entity defines what it considers an acceptable level of risk, often in alignment with its operational objectives, rather than accounting for broader societal concerns. This process-heavy model shifts the focus away from public accountability and onto internal organisational compliance, creating room for external risks to go unchecked.

Conformity Assessments and the Weakness of Self-Certification

The framework's heavy reliance on conformity assessments introduces a significant weakness in the certification process for AI systems. Conformity assessments, while valuable for ensuring that systems meet minimum technical standards, are often internally driven or conducted by third parties closely tied to the organisation. This can lead to potential conflicts of interest and bias in the assessment process. Unlike independent audits, which introduce an external layer of scrutiny, conformity assessments tend to prioritise organisational goals, which may not fully align with societal safety and ethical standards.

The challenge with conformity assessments lies in their assumption that compliance at the time of certification ensures ongoing safety and ethical integrity throughout the lifecycle of the AI system. This is an unrealistic expectation given the adaptive nature of AI, particularly neural networks that continuously evolve. Once a system has been certified, there is little to ensure that its risk profile has not dramatically shifted as it learns from new data and environments. This is especially problematic for general-purpose AI models that can be deployed across multiple, unforeseen use cases.

Moreover, the current process places the burden of compliance documentation and risk evaluation on the deploying organisation, which can lead to a checkbox compliance culture. Organisations may focus on meeting the minimum requirements necessary for certification

without deeply engaging with the broader societal and ethical implications of their AI systems.

Independent audits would provide a more robust and transparent mechanism for ensuring that AI systems continue to operate safely and in alignment with public values throughout their lifecycle. Such audits would also help address the critical concern of public trust, which can be eroded when organisations self-certify compliance without external verification.

Regulatory Approach and the Limitations of Fines as a Deterrent

While the framework for mandating compliance with the proposed guardrails remains under discussion, fines have been proposed as one of the primary enforcement mechanisms for non-compliance. However, when applied to large corporations, the effectiveness of this approach as a genuine deterrent is questionable. Large multinational organisations often view fines as an operational cost rather than a penalty that necessitates behavioural change. The risk, in this context, is that fines do little to promote long-term safety or encourage meaningful shifts in how AI systems are developed, managed, or deployed.

This concern is particularly pronounced in high-risk environments, where the potential for harm is significant. The current proposal appears to allow developers, including those involved in developing high-risk AI systems, to operate even when they are not fully compliant with mandatory standards. Partial compliance, particularly in areas where the consequences of failure are severe, is insufficient. Without stronger deterrents, such as operational restrictions or more stringent regulatory oversight, fines may simply allow organisations to continue operating in ways that do not fully align with the public's interest in safety and accountability.

Given these risks, there should be clear provisions within the framework that allow for the outright banning of developers or AI systems that consistently fail to meet the required safety standards. High-risk developers, especially those whose systems pose direct threats to societal well-being or individual rights, must face consequences that go beyond financial penalties. The capacity to impose bans on such entities would not only serve as a more effective deterrent but also ensure that AI systems operating in high-risk settings are held to the highest possible ethical and safety standards. This would provide the necessary oversight and public assurance that non-compliance will not be tolerated in critical sectors.

As noted in the proposal paper, compounding this complexity is the proliferation of AI agents capable of creating other AI systems, which further complicates the question of accountability. As AI becomes increasingly user-driven and capable of self-generating new

models or applications, the traditional notion of a single, identifiable developer becomes blurred. The issue of who is accountable when an AI system autonomously creates another AI model that is then undertakes harmful actions. A broader discussion around accountability, harm and remedy is needed. The current guardrails do not adequately address this complexity, leaving potential gaps in oversight and responsibility with the prospect of autonomous agents. In a world where AI creation is decentralised, regulatory measures must evolve to track the layers of development and ensure that accountability extends across the entire lifecycle of AI systems, from inception to deployment. Further, the power of AI creates an unprecedented imperative to build a whole of society compliance and safety culture in use, development and engagement.

Absence of a Decommission Phase: Risks in Data Privacy, Ownership, and IP Rights

One glaring gap in the proposed guardrails is the absence of a decommission phase for AI systems. This oversight leaves organisations and regulators without clear guidelines for managing AI systems that have reached the end of their operational lifecycle. Given that AI systems are deeply integrated with sensitive data and proprietary algorithms, the failure to establish decommissioning protocols introduces significant risks around data privacy, intellectual property (IP) rights, and the continued use of outdated models.

Data privacy and ownership are especially pertinent in AI systems that process large amounts of personal information. Without a clear decommissioning process, there is a risk that sensitive data could be retained or misused long after the system is no longer in active use. Additionally, models trained on outdated data could continue to influence decision-making processes if not properly retired. In high-risk sectors such as finance or law enforcement, this creates the possibility of outdated models making critical decisions that no longer align with current standards or data environments.

IP rights also become complex in the absence of decommissioning protocols. The proprietary nature of many AI models means that they carry significant value even after their primary use has ended. Organisations need clear guidelines on how to manage the retirement of these systems to ensure that IP rights are respected and that proprietary knowledge is not inadvertently exposed or misused.

Values Agnosticism and Ethical Minimisation: Public Accountability at Risk

The framework's process-heavy approach sidesteps critical ethical considerations by operating under a values-agnostic model. This is a significant concern, as AI systems are inherently embedded in contexts that require moral scrutiny—whether it's in law enforcement, financial decision-making, or content moderation. By focusing primarily on compliance, the guardrails allow organisations to evade deeper ethical questions about how their AI systems impact individuals, society, and public welfare.

This values agnosticism creates an ethical vacuum where AI systems can be optimised for organisational efficiency, potentially at the expense of societal responsibility. AI systems deployed in high-risk environments—such as those making decisions about creditworthiness, predictive policing, or hiring—are shaping public life in profound ways. Yet, the framework's lack of engagement with these societal risks allows organisations to prioritise internal goals over the broader public good.

Without clear ethical guidelines or a requirement for public accountability, the framework leaves AI systems vulnerable to misuse, manipulation, or the perpetuation of biases. This is particularly concerning given AI's capacity to influence and manipulate behaviour on a large scale. By reinforcing a neutral, process-driven approach, the framework does little to ensure that AI systems contribute to the collective good, leaving room for socially irresponsible applications to emerge unchecked.

Refining the Guardrails for Long-Term Public Accountability

While the proposed mandatory guardrails are an important step toward regulating AI in high-risk settings, several critical gaps need to be addressed. There is an opportunity in opting for free standing legislation to remedy the gaps in the current approach. In particular the reliance on corporate self-assessment of risk categorisation, conformity of compliance obligations, the absence of a decommission phase, and the prioritisation of organisational risk management over public accountability create significant vulnerabilities in the framework.

A more dynamic approach is necessary—one that incorporates independent audits, continuous risk reassessment, and clear ethical guidelines to ensure that AI systems operate safely and responsibly throughout their lifecycle. The framework must also account for the evolving nature of AI systems, particularly neural networks and deep learning models, which do not remain static. Risk cannot be fully determined at the point of deployment, and the

framework should reflect this by requiring ongoing oversight and public accountability. Without these refinements, the guardrails will struggle to keep pace with the very technologies they seek to regulate, leaving societal risks unaddressed.

The Omission of Diversity, Inclusion, and Fairness in AI Governance - Guardrail 10:

One of the most significant oversights in the proposed mandatory guardrails is the exclusion of Guardrail 10, which previously focused on stakeholder engagement, diversity, inclusion, and fairness. In the context of AI governance, particularly in high-risk settings, these considerations are not ancillary—they are fundamental. The absence of Guardrail 10 leaves a substantial gap in the regulatory framework, particularly given the known risks that AI systems can perpetuate or even exacerbate existing biases and social inequalities.

AI systems, by their nature, are not neutral. They are deeply influenced by the data they are trained on and the objectives they are programmed to optimise. Without explicit attention to fairness, diversity, and inclusion, there is a heightened risk that these systems will reinforce existing biases, particularly those related to race, gender, and socioeconomic status. This is especially concerning in areas like predictive policing, hiring algorithms, and financial decision-making, where biased AI systems have the potential to entrench discrimination and inequality.

Bias as an Inherent Feature of AI Systems

All AI systems, particularly those that rely on machine learning and neural networks, reflect the biases present in their training data. Whether the bias is related to race, gender, or socioeconomic status, AI systems can replicate and amplify these patterns, leading to discriminatory outcomes. Without specific guardrails that focus on fairness and inclusion, these biases are unlikely to be corrected. The correction of bias is not just a technical challenge; it is a values-driven decision that requires deliberate engagement with ethical principles.

By omitting Guardrail 10, the framework overlooks the need for AI systems to be assessed not just for technical accuracy but also for their impact on equity and social justice. The correction of bias requires more than just optimising algorithms for technical performance; it requires organisations to engage with questions about the societal impact of their systems and the potential harms they might perpetuate. This is a gap that the current process-driven framework struggles to address, as ethical considerations are not easily quantified or incorporated into traditional compliance models.

The Broader Societal Impact: From Individual to Collective Harms

Guardrail 10 was critical in ensuring that AI systems are designed and deployed with an awareness of their broader societal impacts. The current framework tends to focus on individual-level harms, such as privacy violations or discriminatory decisions affecting specific users. While these are important, they do not fully capture the collective harms that can arise from the widespread deployment of AI systems. AI systems have the capacity to reshape societal structures, influencing public discourse, access to opportunities, and even democratic processes.

For example, AI-driven content curation algorithms on social media platforms can contribute to societal polarisation by reinforcing users' existing beliefs and filtering out opposing viewpoints. Over time, this can undermine social cohesion and contribute to the erosion of public trust in institutions. The exclusion of Guardrail 10 leaves these broader societal risks unaddressed, as the current framework focuses more on technical compliance and individual-level risk than on the collective impact of AI technologies.

Stakeholder Engagement: A Missed Opportunity for Inclusive AI Governance

The exclusion of Guardrail 10 also represents a missed opportunity for inclusive stakeholder engagement in AI governance. Effective AI governance requires input from a broad range of stakeholders, particularly those from marginalised or vulnerable communities who are most at risk of being adversely affected by biased AI systems. Stakeholder engagement is not merely a box to be ticked; it is a critical process for ensuring that AI systems are designed and deployed in ways that are fair, equitable, and aligned with societal values.

Without Guardrail 10, there is little incentive for organisations to engage with these broader stakeholder groups, leading to a governance model that is driven primarily by the interests of the organisations developing and deploying AI systems, rather than by the needs of the communities affected by them. This exclusion undermines the legitimacy of the AI governance framework, as it fails to account for the voices of those most vulnerable to the unintended consequences of AI deployment.

The Need for a Values-Driven Approach

Guardrail 10 provided an essential check on the values embedded in AI systems. By focusing on diversity, inclusion, and fairness, it pushed organisations to engage with the ethical implications of their AI technologies, ensuring that they are not only safe but also

socially responsible. The current framework, by omitting these considerations, runs the risk of allowing AI systems to operate in a values-agnostic manner, where questions of fairness and equity are sidelined in favour of technical optimisation and organisational efficiency.

The omission of Guardrail 10 represents a step backward in the ethical governance of AI. The framework must include clear guidelines that address the societal impact of AI systems, with a particular focus on ensuring that these technologies promote diversity, inclusion, and fairness. Without these principles embedded in the framework, there is a real risk that AI will perpetuate the very inequalities it has the potential to mitigate.

Recommendations for Artificial Intelligence in High-Risk Australian Settings

Under the current Voluntary AI Safety Standards proposed it could be interpreted that a high-risk AI system could still be deployed as long as the developer or organisation could argue that there was no direct intent to cause harm and that the system complies with the ten guardrails. This approach places Australia at significant risk, as it allows potentially harmful AI systems to operate without the robust protections that other jurisdictions, such as the EU, mandate through explicit prohibitions.

The framework's focus is on ensuring that AI systems adhere to broad principles rather than outright prohibiting harmful practices. This means that AI systems designed for manipulative, exploitative, or even highly invasive purposes could still be deployed under the current structure, provided they are developed in a way that appears to meet the guardrails. Essentially, organisations are not required to demonstrate that their AI systems avoid harm altogether; rather, they must only show that they have complied with the guardrails and had no direct intent to cause harm.

This loophole exposes Australia to substantial risks, particularly in high-risk sectors such as healthcare, employment, and criminal justice, where the consequences of biased or discriminatory AI systems can be profound. AI systems in these areas often carry the potential for significant harm, especially to marginalised communities who already face structural barriers. Yet, under the proposed framework, the mere compliance with the guardrails—without addressing the fundamental design flaws or harmful capabilities of these systems—would allow these systems to continue operating.

Furthermore, the framework does not provide sufficient clarity on whether AI systems deployed in high-risk settings, such as diagnostics in healthcare, would always need to fully comply with the guardrails. Diagnostic AI systems, which are used to assess patient conditions and recommend treatments, carry enormous potential for harm if biased or incorrectly trained. However, under the current principles, diagnostic AI systems could theoretically bypass some of the guardrails as long as the developer could argue that the system was developed for its intended use and that there was no direct intent to harm patients. This lack of stringent compliance requirements, particularly in such critical sectors, places patient safety at risk and leaves healthcare providers vulnerable to deploying AI tools that may inadvertently harm marginalised people and communities, such as First Nations people or migrant women, who often face disparities in healthcare access and treatment.

Additionally, the broad wording of what qualifies as high-risk AI leaves much to be desired. While the proposed framework defines high-risk AI as systems that could negatively affect individual rights, health, legal standing, or social wellbeing, it does not provide clear thresholds for when a system qualifies as high-risk. This leaves significant interpretive power in the hands of organisations developing and deploying AI systems, potentially allowing systems with harmful potential to avoid being classified as high-risk altogether. Moreover, even if an AI system is deemed high-risk, the lack of prohibitions on certain harmful practices, such as the collection of predictive personal data or the exploitation of vulnerable groups, means that these systems could still be deployed, provided they comply with the broad, non-specific guardrails.

In contrast, international frameworks, particularly the EU AI Act, take a much stricter stance by explicitly prohibiting certain types of AI systems outright, such as those designed to manipulate or exploit individuals, or those that create social scoring systems based on predictive personal information. These prohibitions ensure that AI systems with a clear potential for harm are not simply permitted to operate after meeting minimal standards of compliance. By not including similar prohibitions, the Australian framework leaves the country at a disadvantage, where potentially discriminatory AI systems could be deployed with little accountability.

The EU AI Act offers a clear and robust framework for defining high-risk AI categorising systems that impact employment, healthcare, education, and law enforcement as inherently high-risk. These sectors are regulated more stringently due to their potential to affect fundamental rights. Australia's current approach, in contrast, lacks this level of precision and may fail to protect intersectional women if high-risk settings are defined too narrowly.

Further, there appears to be a lack of clarity on who will determine which AI systems qualify as high-risk and how these determinations will be made. may slip through the cracks and be subject to weaker oversight. While this may be addressed later in the drafting of legislation, a system of self-reporting would further exacerbate the risks already identified above.

Recommendation:

1. An independent body should be established to evaluate and identify the "high-risk" status of an AI system. This should not be done by self-referral, either by the procurer or developer, but should be applied to each application for use in Australia. To ensure that marginalised groups are adequately protected, the Australian Government should adopt a more precise and inclusive definition of high-risk AI settings, as seen in the EU AI Act. This definition should include sectors like employment, public services, and healthcare, where AI decisions can significantly impact the economic mobility, health outcomes, and social inclusion of vulnerable population. Furthermore, the definition should include an intersectional risk assessment to ensure that the unique needs of migrant women, women of colour, and LGBTIQ+ individuals are addressed.

   It is noted that as the risks of General-Purpose Artificial Intelligence in a globally connected world are difficult to regulate for at a domestic level. As such no recommendations in relation to this are made.

The Evolution of Low-Risk to High-Risk AI: When AI Goes Unmonitored

While the guardrails will apply only to high-risk AI systems, it is essential to acknowledge that AI systems initially classified as low-risk can evolve into high-risk systems over time. This evolution can occur due to unforeseen interactions between AI systems and vulnerable populations, or because of changes in the system's functionality or user base.

Gaps and Concerns: The current proposals for defining and regulating high-risk AI do not account for low-risk AI systems that may evolve into high-risk systems. This gap is particularly concerning for marginalised groups, who may be more vulnerable to the negative impacts of AI systems as they change over time. An AI system that works well in one context

could become discriminatory or harmful when used in different cultural settings or when exposed to biased user inputs.

Recommendation:

2. The Australian Government should adopt a more dynamic approach to AI regulation that allows for reclassification of AI systems as high-risk if they exhibit biases or unintended consequences after deployment. This could include a monitoring process that regularly reviews AI systems for changes in performance, especially when they interact with marginalised populations. The US FDA provides a model for such an approach in its regulation of Software as a Medical Device (SaMD), which requires continuous post-market monitoring to detect emerging risks.

Data Skew, Algorithmic Updates, and the Difficulty in Determining Elevated Risk

Another critical challenge in the regulation of AI systems is the issue of data skew and algorithmic updates, both of which can significantly affect how AI systems perform over time. as the under-diagnosis of certain conditions in people of colour or First Nations people. AI systems frequently undergo independent algorithmic updates, either through the introduction of new data or through changes in the system's design. These updates can inadvertently introduce new biases or amplify existing biases, making it difficult for regulators to determine when an AI system that was initially classified as low-risk has evolved into a high-risk system.

Gaps and Concerns: Australia's current voluntary standards do not provide sufficient guidance on how to manage the risks associated with data skew or algorithmic updates. Without continuous monitoring and bias auditing, AI systems may unintentionally shift from low-risk to high-risk status as their underlying algorithms evolve. This is particularly problematic for AI systems used in healthcare, employment, or public services, where marginalised groups are already at a disadvantage and may face further harm due to biased AI systems.

Recommendation:

3. The Australian Government should mandate that organisations implement ongoing bias audits and algorithmic monitoring to detect changes in risk levels. This approach would help ensure that AI systems that were initially classified as low-risk do not evolve into high-risk systems without proper oversight. In the EU AI Act, organisations are required to perform regular impact assessments and bias audits to ensure that AI systems remain fair and do not introduce discriminatory biases over time.

The Lack of Clarity Over Responsibility for Adverse Outcomes

One of the significant gaps in the proposed voluntary standards and the potential future mandatory guardrails is the lack of clarity around who is responsible for adverse outcomes when an AI system causes harm. AI systems often involve multiple parties across the AI supply chain, including developers, deployers, and end-users. When an AI system makes a biased decision or causes harm—such as a discriminatory hiring decision or a misdiagnosis in healthcare—it is often unclear who should be held accountable.

Gaps and Concerns: The current voluntary standards do not clearly define who is responsible for adverse outcomes, which creates significant legal ambiguity. This lack of clarity is particularly concerning for marginalised groups, who may be disproportionately affected by AI systems in high-risk sectors like employment or law enforcement. If an AI system deployed by a government agency makes a discriminatory decision, who is held accountable: the developer of the AI system or the agency that deploys it?

International Comparison: The EU AI Act provides a clearer framework for accountability, requiring both AI developers and deployers to take responsibility for the outcomes of AI systems. In particular, the EU AI Act mandates that AI systems used in high-risk settings must undergo rigorous testing and must meet compliance requirements that assign clear accountability for any harm caused.

Recommendation:

4. To improve accountability in Australia's AI regulatory framework, the Government should introduce clear guidelines on who bears responsibility for adverse outcomes. This could involve a shared responsibility model, where both developers and deployers are held accountable for ensuring that AI systems do not cause harm, particularly in high-risk sectors. Additionally, Australia should consider adopting an AI ombudsman or regulatory body to investigate cases of AI-related harm and assign responsibility when necessary.

Australia's Role as an Importer and Adopter: Challenges with a Soft Regulatory Approach

Australia is likely to be an importer and adopter of AI technologies, rather than a developer of AI systems at scale. This poses challenges for regulation, particularly under a soft regulatory approach where the focus is on voluntary standards rather than strict enforcement. As Australia imports AI technologies from other jurisdictions, it may face difficulties in controlling the development process of these systems, making it harder to ensure that AI systems meet local standards of fairness, inclusivity, and intersectionality.

Gaps and Concerns: The voluntary nature of the current AI safety standards, combined with Australia's reliance on imported AI technologies, means that Australia may have limited influence over how AI systems are developed, tested, and governed. This creates a significant risk that AI systems deployed in high-risk settings in Australia—such as healthcare or employment—may not be designed with the needs of marginalised Australian women in mind.

International Comparison: In contrast, the EU AI Act imposes stringent requirements on imported AI systems, ensuring that any AI technology deployed within the EU meets the same high standards for bias auditing and transparency as AI systems developed domestically. The EU's strong regulatory framework allows it to exert greater control over imported AI technologies, ensuring that they align with the EU's ethical standards.

Recommendation:

5. To address the challenges of being an importer and adopter of AI technologies, Australia should implement stronger regulatory mechanisms that ensure imported AI systems meet the same ethical and safety standards as domestically developed systems. This could include mandatory pre-market testing and compliance checks for all AI systems deployed in high-risk settings, regardless of where they were developed. Additionally, Australia should consider adopting a whole-of-economy approach, as outlined in Option 3 of the consultation paper, to ensure that AI regulation is consistent and enforceable across all sectors.

This expanded section addresses the voluntary nature of the current standards, the challenges of defining and managing high-risk AI, and the potential for low-risk systems to evolve into high-risk systems, particularly when interacting with marginalised groups. It also highlights the complexities of accountability in the AI supply chain and the difficulties Australia may face as an importer and adopter of AI technologies.

How to Strengthen the Regulatory Approach

Given the disruptive and profound impact that AI is likely to have across all sectors of the economy, it is crucial that a comprehensive and multi-faceted regulatory approach is adopted to effectively manage and mitigate the risks associated with its deployment. AI has the potential to revolutionise industries ranging from healthcare to financial services, yet this transformative power also introduces significant risks, particularly in high-risk sectors where errors or biases in AI systems could result in serious harm to individuals and communities. Therefore, it is recommended that the Australian Government pursue option three—the introduction of a new cross-economy AI-specific Act, such as an Australian AI Act. This would establish a clear legal framework dedicated to the governance and regulation of AI technologies across all sectors, ensuring that ethical standards, safety measures, and accountability structures are codified in a unified legislative approach.

However, while an economy-wide framework is essential, it will also be necessary to develop sector-specific frameworks that can address the unique challenges and considerations posed by AI in particular industries. For instance, the use of AI in healthcare introduces specific ethical concerns, such as patient privacy, data protection, and the risk of biased algorithms influencing clinical decision-making. Therefore, a sector-specific AI framework for healthcare should be developed to provide additional safeguards tailored to the intricacies of that field, ensuring that AI deployment improves healthcare outcomes without compromising patient rights or safety. This sectoral approach should also extend to other critical industries, including education, law enforcement, and financial services, where AI could significantly influence public trust, access to services, and equity.

In addition to legal frameworks, it is also recommended that a carefully designed regulatory ecosystem be created to ensure ongoing oversight, risk assessment, and accountability in AI deployment. Central to this ecosystem would be the establishment of an independent body tasked with assessing the risks associated with AI systems, both before and after deployment. This body would play a crucial role in evaluating whether AI technologies meet ethical and safety standards, ensuring that AI systems do not disproportionately harm vulnerable populations or undermine public confidence. Further, this independent entity could serve as a central point of oversight, facilitating transparency and public engagement in AI governance.

Complementing this independent body, the creation of an independent auditing sector specifically focused on AI systems would be critical for ensuring that AI technologies are regularly reviewed and monitored for compliance with both general and sector-specific standards. These independent auditors would be responsible for assessing the performance, accuracy, and fairness of AI systems, particularly those used in high-risk sectors like healthcare and financial services, where the potential for harm is particularly high. Such an auditing sector would not only help to identify and address issues related to algorithmic bias or data integrity but would also provide an essential layer of accountability by ensuring that AI systems are continually evaluated and improved based on real-world outcomes.

Together, the introduction of a cross-economy AI Act, the development of sector-specific frameworks, and the creation of an independent regulatory ecosystem would establish a robust and comprehensive approach to AI governance in Australia. This would provide a clear path forward for responsible AI development, mitigating risks while promoting innovation and ensuring that the benefits of AI are distributed fairly across society. Without such measures, Australia risks falling behind international best practices in AI governance,

leaving vulnerable groups and critical sectors exposed to the unregulated deployment of powerful and potentially harmful AI technologies.

Auditing AI at the Development Stage

The introduction of AI technologies requires a rigorous approach to safety, fairness, and accountability, particularly in healthcare where the consequences of failure can be severe. While there are well-established frameworks for AI ethics and safety, the challenge is to turn these high-level principles into actionable processes. This involves embedding safety measures across the lifecycle of AI development, deployment and use, ensuring comprehensive assessment and review processes, and building an ecosystem that is adaptable to the fast-evolving nature of AI technologies.

Currently, Australia's voluntary AI guardrails focus primarily on the deployment of AI systems, with limited attention to auditing and oversight during the development stage. However, many of the most harmful biases in AI systems arise at this stage, when the system's architecture, datasets, and algorithms are first designed. Without rigorous development-stage auditing, organisations may unknowingly introduce biases that disproportionately affect marginalised people and communities, making it difficult to address these biases later in the system's lifecycle.

Incorporating lifecycle awareness is crucial to ensuring AI systems meet safety standards not just at deployment but continuously as they evolve. AI systems must undergo continuous auditing and monitoring to adapt to the changing nature of data inputs, healthcare demands, and technological advances. Beyond operational audits, a comprehensive AI safety approach should address social responsibilities. These systems should not only be safe for patients but also ethical in their societal impact, helping foster public trust.

Building public trust is essential, and the role of independent auditing becomes even more critical as it ensures AI systems are evaluated regularly to meet ethical, fairness, and safety standards. This is particularly important as AI systems continue to evolve after deployment, maintaining trust in the system over time.

Gaps and Concerns: The absence of early-stage auditing is a major gap in Australia's AI governance framework. AI systems designed with biased data or algorithms are likely to produce discriminatory outcomes that will only become evident after deployment, by which point the harm may be difficult to reverse. For example, AI systems used in healthcare diagnostics or public services may be trained on datasets that underrepresent First Nations

people, migrants, or individuals from low-income communities, leading to biased recommendations or decisions that exacerbate health and social inequalities.

International Comparison: In contrast, the EU AI Act requires comprehensive pre-market assessments for high-risk AI systems, including audits at the development stage to identify potential biases, safety risks, and transparency issues. These early-stage audits ensure that AI systems are designed to meet safety and fairness standards before they are deployed in critical sectors like healthcare, employment, and law enforcement.

Recommendation:

6. Australia should mandate auditing at the development stage, requiring AI developers to conduct thorough reviews of their datasets, algorithms, and system architectures to identify and mitigate intersectional biases. These audits should focus on ensuring that AI systems are designed to be fair, inclusive, and safe for marginalised communities, particularly women of colour, LGBTIQ+ individuals, and women with disabilities. Additionally, auditing should assess whether AI systems are trained on diverse datasets that accurately reflect the experiences and needs of vulnerable populations.

Assessing Safety Culture in AI Development

Another essential component of building a regulatory ecosystem is the cultivation of a safety culture within organisations that develop and deploy AI systems. A strong safety culture ensures that ethical considerations, risk mitigation, and inclusivity are prioritised at every stage of AI development and deployment. However, Australia's current voluntary AI standards do not require organisations to assess their internal safety culture or to ensure that AI systems are developed with a focus on protecting vulnerable groups.

Gaps and Concerns: A weak safety culture within AI development teams can lead to the creation of systems that do not adequately account for the risks faced by marginalised people and communities. For example, if developers are not trained to recognise how biases can emerge in AI systems, they may overlook intersectional risks, such as how an AI system used in employment could disproportionately disadvantage migrant women or how a healthcare AI system could fail to diagnose conditions that disproportionately affect women of colour.

International Comparison: The US FDA guidelines for AI in medical devices include a strong emphasis on safety culture, requiring organisations to integrate risk management practices and ensure that ethical considerations are prioritised throughout the system's lifecycle. These guidelines focus on the importance of creating a culture of safety that permeates all levels of AI development, from initial system design to post-market monitoring.

Recommendation:

7. Australia should incorporate requirements for organisations to assess and build a safety culture within their AI development teams. This could include mandatory training programs on AI ethics, intersectionality, and bias mitigation, as well as the creation of internal safety committees that oversee the ethical development of AI systems. Additionally, organisations should be required to submit safety culture reports as part of their compliance with AI safety standards, ensuring that ethical considerations are embedded in the organisational culture from the outset.

Training, Upskilling, and Mitigating the Risks of Skill Attrition

As AI systems become more integrated into high-risk sectors such as healthcare, science, engineering, and public services, there is growing concern about the risk of enfeeblement—the loss of critical human skills due to over-reliance on AI technologies. This is particularly concerning in fields where human expertise is vital for making complex, ethical decisions, such as in medical diagnostics or scientific research. Additionally, the rapid adoption of AI technologies may lead to skill attrition in critical sectors, leaving workers—particularly women—vulnerable to job displacement or skill degradation.

Enfeeblement and the Loss of Critical Skills

One of the primary risks associated with AI integration is enfeeblement, the phenomenon whereby human workers lose critical skills as they become overly reliant on AI systems to perform tasks. This is especially concerning in sectors such as healthcare and engineering, where human judgment and expertise are essential for ensuring safety, ethical decision-making, and complex problem-solving.

Gaps and Concerns: Without proper safeguards, the increasing reliance on AI in high-risk sectors could lead to the erosion of critical human infrastructure, such as medical professionals, engineers, and scientists. Women working in these fields—particularly those

in junior or mid-level positions—may be disproportionately affected, as they are more likely to face job displacement or skill attrition due to AI systems taking over routine tasks. For example, if AI systems are used to automate routine diagnostic tasks in healthcare, junior medical professionals may lose opportunities to develop essential diagnostic skills, leaving them less prepared for complex decision-making in the future.

International Comparison: In the UK's National Health Service (NHS), there are ongoing discussions about how to integrate AI technologies into healthcare without leading to the loss of critical human skills. The NHS is investing in training programs to ensure that healthcare professionals remain engaged in clinical decision-making while using AI tools to enhance their abilities rather than replace them.

Recommendation:

8. Australia should implement policies to mitigate the risks of enfeeblement by requiring that AI systems in high-risk sectors be used to augment, rather than replace, human decision-making. This can be achieved by creating hybrid models in which human professionals work alongside AI systems, ensuring that they continue to develop and maintain critical skills. Additionally, training programs should be established to help professionals in healthcare, engineering, and science understand how to use AI tools effectively while maintaining their core competencies.

Upskilling and Reskilling Workers to Adapt to AI Technologies

As AI continues to transform industries, there is an urgent need to provide upskilling and reskilling opportunities for workers, particularly for women in sectors where automation is leading to job displacement. Many women—especially those in marginalised groups—may face barriers to accessing these training opportunities, exacerbating existing inequalities and limiting their ability to adapt to the changing job market.

Gaps and Concerns: The current AI safety standards do not include provisions for upskilling or reskilling workers affected by AI technologies. This is particularly concerning for women in vulnerable sectors, such as telecommunications, public services, and customer service, where AI systems are increasingly being used to automate routine tasks. Without access to

training and support, many of these women may be left behind in the transition to an AI-driven economy.

International Comparison: In Singapore, the government has launched the SkillsFuture initiative, which provides funding for workers to undergo reskilling and upskilling programs in emerging technologies, including AI. This initiative is aimed at ensuring that workers remain competitive in the job market as industries evolve due to technological advancements.

Recommendation:

9. Australia should develop a national strategy for upskilling and reskilling workers, with a particular focus on ensuring that women—especially those from marginalised groups—have access to the resources they need to adapt to AI technologies. This could include government-funded training programs, incentives for organisations to offer upskilling opportunities, and partnerships with educational institutions to create AI literacy programs. Additionally, specific provisions should be made for intersectional women, ensuring that training programs are accessible to migrants First Nations people, and other marginalised groups who may face additional barriers to education and employment.

This extended analysis addresses the gaps in Australia's AI approach concerning the regulatory ecosystem, including auditing at the development stage, safety culture, and the need for upskilling and reskilling to prevent the loss of critical human skills.

An Example of Integrating AI Guardrails with Existing Systems and Frameworks - AI and Healthcare Equity

Artificial Intelligence (AI) is reshaping healthcare globally, offering new possibilities in diagnosis, treatment personalisation, and access to care. In Australia, AI is being integrated into healthcare systems through various initiatives, including by the Department of Health and Aged Care, and the Department of Industry, Science and Resources. However, while these technologies promise significant advancements, there are ethical challenges that must be addressed—particularly regarding marginalised populations such as migrants and

refugees especially women, LGBTIQ+ individuals, Aboriginal and Torres Strait Islander people, people with disability and vulnerable women. To ensure AI serves all Australians equitably, it is essential that its implementation does not replicate systemic biases or create new barriers.

AI in Healthcare: Potential Risks and Ethical Challenges

AI will increasingly be integrated into healthcare settings, offering potential solutions to improve diagnostic accuracy, streamline clinical workflows, and even expand healthcare access in rural or remote areas. AI's ability to process vast amounts of data and identify patterns has already demonstrated its capacity to assist in areas like radiology, oncology, and pathology, leading to faster diagnoses and more personalised treatments. Despite these promises, both healthcare professionals and consumers are raising significant concerns about the ethical and practical implications of AI's integration into healthcare, particularly around bias, privacy, patient autonomy, and the possibility skills might atrophy.

A significant issue raised by health providers and the public is the potential for AI to perpetuate existing biases within the healthcare system. AI systems are often trained on historical data, which can reflect societal inequalities and entrenched healthcare disparities. Australia's diverse population—including migrants and refugees, Aboriginal and Torres Strait Islander people, people with disability and second- generation culturally and linguistically diverse (CALD) groups—AI systems risk producing inaccurate diagnoses or treatment recommendations if they are trained primarily on data from urban, English-speaking, and often white populations. For example, an AI trained on such data may underperform when diagnosing diseases in people with darker skin tones or in those who present symptoms differently due to cultural or biological diversity. As a result, vulnerable populations may receive suboptimal care, exacerbating already existing health disparities.

Surveys of healthcare consumers in Australia reflect these concerns. According to the Australian Alliance for AI in Healthcare (AAAiH), privacy and data security were among the top concerns for both patients and clinicians when it came to AI in healthcare. More than 15% of respondents in a 2021 national survey expressed worries about how their health data would be used, with many fearing that the introduction of AI could lead to misuse or inadequate protection of sensitive information. Vulnerable populations, such as refugees or migrants who may already distrust government or institutional systems, may be particularly hesitant to engage with AI-based healthcare if they believe personal or group data could be misused or exposed to privacy and security risks.

In addition to data privacy, patient autonomy is a crucial issue. AI systems, by their nature, are often seen as “black boxes,” where the decision-making process is opaque even to healthcare providers. This can lead to situations where clinicians feel they are losing control over treatment decisions or are unable to fully explain AI- generated recommendations to patients. Trust is a fundamental element of the clinician-patient relationship, and if patients feel that decisions about their health are being outsourced to an algorithm, this could undermine their confidence in the healthcare system. A survey conducted by the Australian Medical Association (AMA) highlights this point, with clinicians expressing significant concerns about the transparency and explainability of AI tools used in clinical settings. They warned that systems not designed specifically for medical purposes—such as non-medical grade AI—could pose risks to patients if adopted without rigorous testing and validation.

From a regulatory perspective, Australia’s health system is currently unprepared to manage the rapid development and integration of AI technologies in clinical settings. Experts have pointed out that while Australia’s digital infrastructure and healthcare system provide a strong foundation for the implementation of AI, there are substantial gaps in the regulatory framework, workforce training, and governance structures needed to ensure the safe, effective, and ethical use of AI in healthcare. For example, many AI systems currently in use are not subject to the same rigorous testing as medical devices. As a result, non-medical grade AI applications, such as some AI chatbots or virtual assistants, could find their way into clinical use without appropriate safeguards, posing risks to patient safety.

The Australian Government and health bodies have begun making progress in addressing these concerns. The Therapeutic Goods Administration (TGA) currently regulates medical-grade AI, ensuring that systems designed for clinical use meet certain safety and efficacy standards. However, there is growing recognition that more comprehensive governance is required to cover the full spectrum of AI applications in healthcare, particularly those that are not yet classified as medical devices. The AMA has called for a national strategy that includes better regulation, greater transparency in algorithm development, and enhanced data privacy protections. There are also recommendations for more robust education and training programs to ensure that both clinicians and consumers are equipped to understand and engage with AI-based healthcare systems in a way that preserves patient autonomy and trust.

Another significant barrier to the successful implementation of AI in healthcare is workforce capability. A 2023 roadmap developed by the AAAiH identified workforce development as a key priority, noting that many healthcare providers currently lack the skills to critically evaluate AI tools or integrate them effectively into their practice. Both clinicians and patients

need to be educated on the limitations and appropriate uses of AI in healthcare. Without sufficient training, there is a risk that clinicians may over-rely on AI systems or fail to question their outputs, which could lead to patient harm.

At this stage, it is clear that both clinicians and consumers have substantial concerns about its ethical use. Bias in AI systems, data privacy risks, and the erosion of patient autonomy are critical challenges that need to be addressed before AI can be fully integrated into healthcare.

AI Guardrails are Insufficient in a Health Setting

The deployment AI in healthcare has accelerated in recent years, offering new solutions for diagnostics, personalised treatment, and clinical workflows. However, these technological advancements come with significant ethical and practical challenges. As AI systems increasingly influence clinical decision- making and patient care, there is a pressing need for robust frameworks to guide their development, implementation, and oversight. Internationally across governments, international organisations and civil society, an abundance of high-level policy documents on AI ethics, safety and responsibility have been published.

Australia's Voluntary AI Safety Standard announced in Sept 2024, provides a weak framework for AI adoption with a focus is on deployers and secondary guidance on procurement, focusing on governance, risk management, and human oversight. The standards map to the Australian AI Ethics Principles (see below) and are framed as a precursor to the development of future mandatory regulations for high-risk AI systems. Unlike most developed countries, Australia has favoured an innovation narrative with less emphasis on risk mitigation. It does acknowledge the need for care in high-risk areas like healthcare, however a health specific framework is yet to be developed and would be necessary given the associated risks. The key safety mechanism is an emphasis on the retention of human decision-making.

In the current iteration, the proposed mandatory standards also appear to place the major responsibility for regulatory compliance on the deployers of AI systems, posing challenges. While it offers clear guidelines on the need for governance and transparency processes, there is no specific requirements or detailed provisions around duty of care on intersectionality and data bias.

Compared to international approaches, Australia's standards lag behind the more prescriptive frameworks seen in the EU and the U.S. The EU AI Act and the FDA's Software as a Medical Device (SaMD) regulations provide more binding rules for high-risk AI, with

strict accountability and oversight measures. However, while Australia does align with aspects of global standards like ISO/IEC 42001:2023, ensuring international compatibility, the limited nature of guardrails may reduce the effectiveness of mandatory standards, especially in sectors where stricter compliance is required. There is also little indication of how organisations procuring AI should work to ensure compliance.

As it currently stands Australia's framework prioritises voluntary compliance initially, while preparing for mandatory enforcement as public consultation continues. Currently the approach taken to AI safety would not be considered rigorous and would be perceived as very weak next to mandatory AI regulations already in place internationally.

Australia's Artificial Intelligence Ethics Principles

Previously Australia had developed Artificial Intelligence Ethics Principles, drafted by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in 2019. Following this a subsequent consultation process on Implementing Australia's AI Ethics Principles in 2023 with the Department of Industry, Science and Resources, provides a comprehensive foundation for responsible AI integration across all sectors, including healthcare. The framework comprises eight key principles: human-centred values, fairness, privacy, reliability, accountability, transparency, inclusivity, and environmental sustainability. These principles are designed to ensure that AI technologies are developed and deployed in ways that benefit society while mitigating risks. In the context of healthcare, the framework offers a set of guidelines to protect patient rights, improve care quality, and maintain trust in AI-driven healthcare innovations.

The human-centred values principle is particularly relevant in healthcare, where patients' rights to dignity, autonomy, and informed consent are central to ethical medical practice. AI systems, under this framework, are required to respect these values by supporting clinicians in their decision-making processes rather than replacing human judgment. For example, diagnostic AI tools should act as decision support systems, providing clinicians with additional insights rather than making decisions autonomously. This ensures that the final responsibility for patient care remains with human healthcare providers, thus maintaining the central role of human expertise in healthcare.

Similarly, the principles of accountability and transparency are crucial in healthcare, where the consequences of errors or biased decisions can be severe. The framework emphasises the importance of clear accountability mechanisms, ensuring that both AI developers and healthcare providers are responsible for the outcomes of AI-assisted decision-making. This

is particularly important in clinical settings where the use of AI may blur the lines of responsibility. For instance, if an AI tool provides a flawed diagnosis, it is essential that there are mechanisms to determine whether the error lies with the AI developer (for faulty algorithm design) or the clinician (for misinterpreting AI outputs).

While the Artificial Intelligence: Australia's Ethics Framework was in some ways more comprehensive than the new guardrails, they still lacked specific guidelines for addressing the cultural and linguistic diversity of Australia's population. While the framework's inclusivity principle encourages the development of AI systems that do not marginalise vulnerable groups, it does not provide concrete strategies for designing AI tools that are culturally and linguistically competent. In a healthcare system that serves a highly diverse population, including Aboriginal and Torres Strait Islander people, migrants and refugees, and people with disability, this omission is significant. For instance, AI diagnostic tools may fail to account for cultural differences in symptom reporting, or they may not be equipped to provide services in languages other than English. This lack of cultural competence in AI design could lead to misdiagnoses or suboptimal care for patients from CALD backgrounds.

Furthermore, the framework's fairness principle is underdeveloped in terms of operationalising equity in healthcare. Healthcare AI systems often rely on large datasets that reflect historical biases, such as underrepresentation of certain demographic groups in clinical trials or healthcare records. Without specific guidelines on how to identify and mitigate these biases, the fairness principle risks being more aspirational than actionable. For example, AI tools for predicting patient outcomes may disproportionately disadvantage minority groups if they are trained on data that does not adequately represent these populations. In this context, the ethical framework should include more detailed provisions for bias auditing and the development of equity-oriented algorithms.

World Health Organization (WHO) Ethical Guidelines for AI in Healthcare

On the international stage, the World Health Organization (WHO) has developed its own ethical guidance document for the use of AI in healthcare. These guidelines emphasise six key principles: protecting human autonomy, promoting human well- being, ensuring transparency and explainability, fostering inclusivity and equity, ensuring data privacy and security, and promoting accountability. The WHO guidelines aim to create global standards for the ethical use of AI in healthcare, ensuring that AI technologies contribute to better health outcomes without exacerbating existing inequalities.

One of the central tenets of the WHO guidelines is the protection of human autonomy. This principle is particularly relevant in healthcare, where patient autonomy is a foundational ethical concern. The WHO guidelines stress that AI systems should support, rather than replace, human decision-making in clinical settings. This approach aligns with traditional ethical norms in medicine, where clinicians are expected to respect patients' rights to make informed decisions about their own care. In practice, this means that AI tools should be designed to enhance clinical judgment by providing supplementary information, rather than making autonomous decisions about patient care.

The WHO guidelines also place a strong emphasis on transparency and explainability. In healthcare, it is essential that both clinicians and patients understand how AI systems arrive at certain decisions or recommendations. The black-box nature of many AI algorithms, particularly those based on deep learning, poses a challenge to achieving this transparency. If AI systems are not transparent, clinicians may struggle to interpret the recommendations, and patients may lose trust in AI-driven care. The WHO guidelines advocate for the development of AI systems that are explainable, meaning that the reasoning behind their outputs can be easily understood by users. This is particularly important in healthcare, where the stakes are high and trust between clinicians and patients is paramount.

Despite these strengths, the WHO guidelines face similar limitations to Australia's ethics framework when it comes to addressing the needs of vulnerable populations. While the guidelines advocate for inclusivity and equity, they do not provide specific recommendations on how AI systems can be designed to serve marginalised groups, such as migrants, refugees, or those from CALD backgrounds. The guidelines acknowledge the risk of bias in AI systems but stop short of providing detailed strategies for identifying and mitigating these biases in healthcare settings. As a result, there is a gap between the high-level ethical principles outlined by the WHO and the practical steps needed to ensure that AI systems are truly inclusive and equitable.

For example, AI systems used in global health initiatives may fail to account for the unique healthcare challenges faced by refugees, who often experience barriers to accessing healthcare due to legal, linguistic, and cultural factors. Without specific guidance on how to address these challenges, the inclusivity and equity principles may not be fully realised in practice. Additionally, the guidelines do not adequately address the issue of language accessibility, which is critical in healthcare settings where patients may not speak the dominant language of the healthcare provider. AI systems that rely on natural language processing (NLP) technologies must be designed to support multiple languages and dialects to ensure that all patients can benefit from AI-driven healthcare innovations.

UNESCO Recommendation on the Ethics of Artificial Intelligence

Another key international framework is the UNESCO Recommendation on the Ethics of Artificial Intelligence, which was adopted in 2021. This framework is notable for its focus on human rights, social justice, and sustainability in the development and deployment of AI technologies. Like the WHO guidelines, the UNESCO framework emphasises the importance of protecting human autonomy, ensuring transparency, promoting inclusivity, and safeguarding data privacy. However, the UNESCO framework goes further by explicitly linking AI ethics to broader social justice concerns, such as reducing global inequalities and ensuring that AI technologies do not disproportionately harm marginalised populations.

One of the strengths of the UNESCO framework is its explicit focus on social justice. In the context of healthcare, this principle aligns with the goal of ensuring that AI technologies are developed and deployed in ways that reduce health disparities, rather than exacerbating them. The framework calls for AI systems to be designed with the specific needs of marginalised groups in mind, including migrants, refugees, and First Nations populations. This focus on social justice is particularly relevant in healthcare, where systemic inequalities often result in poorer health outcomes for vulnerable populations. For example, AI systems used in public health initiatives could be designed to address the unique healthcare challenges faced by refugees, such as access to mental health services or the management of chronic diseases in low-resource settings.

Another key strength of the UNESCO framework is its emphasis on sustainability. While this principle is often applied to environmental concerns, it also has implications for healthcare. AI systems that are designed to promote long-term health equity must be sustainable, meaning that they are accessible, affordable, and adaptable to the needs of diverse populations. For example, AI tools used in rural healthcare settings must be designed to function in environments with limited infrastructure, such as areas with poor internet connectivity or shortages of healthcare professionals. By prioritising sustainability, the UNESCO framework ensures that AI technologies contribute to the long-term resilience of healthcare systems, particularly in underserved communities.

However, like the WHO guidelines, the UNESCO framework faces challenges in terms of operationalising its ethical principles. While the framework provides a strong ethical foundation for the development of AI technologies, it does not offer detailed guidance on how these principles can be translated into practical actions in healthcare settings. For instance, while the framework advocates for inclusivity and equity, it does not provide specific recommendations on how to design AI systems that are culturally and linguistically

competent. This is a significant gap, particularly in multicultural societies like Australia, where healthcare systems must cater to the needs of a highly diverse population.

SPG's Cultural Competency Standards could be built upon to address some of the gaps in both national and international AI ethical frameworks, particularly when it comes to ensuring that healthcare services and tools are tailored to the needs of culturally and linguistically diverse (CALD) populations. These standards emphasise that healthcare systems must be designed to be accessible, culturally sensitive, and responsive to the needs of diverse patient populations, including migrants and refugees. This focus on cultural competency is vital in a multicultural society like Australia, where a significant portion of the population may face language barriers or experience healthcare differently due to cultural beliefs and practices.

The Cultural Competency Standards advocate for the inclusion of CALD voices in the development and implementation of healthcare policies and tools. For AI systems in healthcare, this means that the design process must include input from diverse communities to ensure that the tools are relevant and effective for all patients. For example, an AI system that assists with mental health diagnoses should be trained on diverse data sets that include input from patients of various cultural backgrounds, ensuring that cultural differences in the expression and understanding of mental health issues are accounted for. Similarly, AI systems that provide treatment recommendations should consider cultural factors that may influence a patient's willingness or ability to comply with certain treatment plans.

Moreover, the Cultural Competency Standards stress the importance of language accessibility, which is a critical gap in most existing AI frameworks. Many AI-driven healthcare tools, such as telehealth platforms and diagnostic applications, are predominantly designed for English-speaking users. However, in Australia, where a significant proportion of the population speaks languages other than English at home, this lack of linguistic inclusivity could lead to unequal access to care. AI tools must be equipped with multilingual capabilities, allowing patients to interact with healthcare providers in their preferred language and ensuring that important medical information is conveyed accurately. By integrating these standards into the design and deployment of AI systems, Australia can ensure that AI technologies are inclusive and accessible to all individuals, regardless of their cultural or linguistic background.

OECD Collective Action for Responsible AI in Health

The OECD Collective Action for Responsible AI in Health is a comprehensive policy document that addresses the ethical integration of AI into healthcare systems, focusing on balancing the innovative potential of AI with its ethical challenges. It stresses that AI in healthcare must be deployed in ways that benefit society, healthcare institutions, and most importantly, patients. The document is built on principles that align closely with both AI ethics and health ethics, ensuring that AI advancements do not compromise the foundational values of healthcare. The primary objective of the guidance is to control that AI serves to improve healthcare delivery while upholding standards such as fairness, transparency, and accountability.

The document emphasises inclusive and equitable use to bridge the gap between health ethics and digital inclusion. In particular, it states that AI should be implemented to enhance access to healthcare, particularly for underserved populations. This principle reflects the long-standing health ethic of equitable access to care, ensuring that vulnerable groups are not left behind in the AI revolution.

The OECD guidance stipulates that AI must reduce healthcare disparities and provide broader, more efficient access to medical services, particularly for populations that traditionally face barriers such as geographic location, socio- economic status, or linguistic differences. This focus aligns well with the ethical obligation in healthcare to provide equitable care to all patients, regardless of their background. While the focus is closing the gap between access to health in developed verses developing countries, much of the ethical approach is broadly applicable.

Of particular note, in the application of an intersectional lens is the emphasis the OECD guidance places on human-centred approaches, ensuring that AI systems respect human dignity and autonomy, core values of medical ethics. A key pillar that should be considered in the implementation of AI within the Australian health system is the principle that such systems should support, rather than replace, human decision-making, with healthcare professionals maintaining control over clinical judgments and patient care.

By embedding these safeguards into AI systems, it integrates both AI ethics (which emphasizes non-maleficence and transparency) and the longstanding medical principles of respect for patient autonomy. Further, it outlines the importance of transparency in AI systems, ensuring that healthcare providers and patients understand how decisions are made by AI, allowing for more informed consent and a greater sense of trust in the technology.

The document also addresses the critical aspect of transparency and accountability. In healthcare, it is essential that AI systems can be trusted by both clinicians and patients. The document recommends that AI systems be designed to allow for explainability, ensuring that healthcare providers can understand and validate the recommendations made by AI algorithms.

Additionally, it outlines the need for clear accountability, ensuring that healthcare institutions and AI developers can be held responsible for the outcomes of AI- powered systems. This is particularly important in clinical settings where the consequences of an AI system's failure or bias could have severe implications for patient health.

Given the sensitive nature of healthcare, the OECD advises that AI systems undergo extensive evaluation before they are deployed in real-world medical contexts. This involves ensuring that AI is robust enough to handle complex medical conditions and diverse patient data, a requirement closely related to the ethical imperative in healthcare to "do no harm." The document's attention to security also ties into the ethical need for data privacy, especially in healthcare where patient data is highly sensitive.

Australian Alliance for Artificial Intelligence in Healthcare (AAAiH) Roadmap

The AAAiH National Policy Roadmap in 2023 provides a strategic plan for integrating AI into Australian healthcare responsibly. It focuses on cross-sector and international collaboration, while also supporting healthcare professionals in AI literacy.

However, there are some deficiencies, particularly regarding vulnerable populations. The roadmap could offer more explicit guidelines on addressing the needs of marginalised communities, such as migrants and refugees, women of colour, people who identify as LGBTIQ+ and people with disability. It acknowledges issues like algorithmic bias but could include more measures to combat these biases effectively, particularly for those facing multiple, overlapping barriers. However, it does mention of equity as a guiding principle. The roadmap's focus on creating a continuous feedback loop for a recommendation for auditing AI systems aligns with emerging consensus.

Gaps in Current AI Frameworks for Vulnerable Populations

While the Artificial Intelligence: Australia's Ethics Framework, the WHO's Ethical Guidelines, and UNESCO's Recommendations provide comprehensive guidance on the ethical use of AI, they often fail to adequately address the unique needs of vulnerable populations, such as

migrants and refugees, Aboriginal and Torres Strait Islander people, people with disability and people who identify as LGBTIQ+. One of the primary gaps in these frameworks is the lack of detailed guidance on cultural and linguistic competency, which is essential for ensuring that AI systems in healthcare serve all individuals equitably. As highlighted by the SPG/MRHP Cultural Competency Standards, healthcare services and technologies must be designed to meet the diverse needs of Australia's population, but many AI frameworks do not provide sufficient strategies for achieving this.

A related issue is the lack of representation in AI training data. AI systems in healthcare rely heavily on large datasets to make predictions and recommendations, but if these datasets are not representative of the broader population, the AI's outputs may be biased. This is particularly problematic for vulnerable populations, who may be underrepresented in healthcare datasets. For example, an AI system trained primarily on data from urban, English-speaking, white patients may not perform as well when used to diagnose conditions in First Nations Australians, migrants, or refugees, who may present symptoms differently or have unique health concerns. To address this gap, AI developers must ensure that the data used to train healthcare AI systems is representative of the entire population, including those from diverse cultural, linguistic, and socioeconomic backgrounds.

Another significant gap in current frameworks is the failure to address the power dynamics inherent in the use of AI in healthcare. AI systems are often viewed as authoritative, and there is a risk that patients from vulnerable populations may feel disempowered when interacting with AI-driven healthcare tools. This is particularly concerning for migrants and refugees, who may already have a heightened sense of vulnerability when accessing healthcare due to past experiences of discrimination or systemic exclusion. Ethical AI frameworks must therefore include provisions that ensure patient empowerment and informed consent. This can be achieved by designing AI systems that are transparent and explainable, allowing patients to understand how decisions are being made and ensuring that they retain control over their healthcare.

Regulatory Challenges and the Role of Government

In addition to the ethical gaps identified in existing frameworks, there are also significant regulatory challenges that must be addressed to ensure the safe and effective use of AI in healthcare. In Australia, the Therapeutic Goods Administration (TGA) is responsible for regulating medical devices, including AI systems used in healthcare. While the TGA provides a regulatory pathway for AI technologies, many AI applications in healthcare do not fall under its purview. For example, non-medical- grade AI tools, such as AI-driven chatbots used for

mental health support, are not subject to the same rigorous testing and oversight as traditional medical devices. This regulatory gap poses risks to patient safety, as unregulated AI tools may produce inaccurate or harmful recommendations.

To address this issue, there is a growing call for the Australian Government to develop a more comprehensive regulatory framework for AI in healthcare. This framework should include provisions for the ongoing monitoring and auditing of AI systems to ensure that they perform safely and equitably in clinical practice. Moreover, the framework should mandate the inclusion of bias auditing and equity assessments as part of the approval process for AI healthcare tools, to ensure that these technologies do not disproportionately harm vulnerable populations. By developing a more robust regulatory framework, the Australian Government can help build public trust in AI technologies and ensure that they are used in ways that promote health equity.

Furthermore, the Government has a critical role to play in capacity building for the healthcare workforce. As AI becomes more integrated into clinical practice, healthcare providers must be equipped with the knowledge and skills to use these tools effectively and ethically. This includes training healthcare professionals to recognize the limitations of AI, interpret AI-generated recommendations critically, and communicate AI-driven insights to patients in a way that respects their autonomy and cultural background. The Australian Department of Health and Aged Care and the Department of Industry, Science and Resources can facilitate this by partnering with universities, healthcare institutions, and AI developers to offer training programs and resources for clinicians.

Regulatory approaches in other international jurisdictions

The European Union has implemented a robust approach to AI safety in healthcare, primarily through the Artificial Intelligence Act and the Ethics Guidelines for Trustworthy AI. The AI Act classifies AI systems based on risk, with healthcare AI falling into the high-risk category due to its direct impact on patient health. These regulations ensure that AI systems in healthcare undergo strict testing and approval processes before being implemented. The focus is on transparency, requiring AI systems to be explainable to both healthcare providers and patients. Additionally, the AI systems must comply with the General Data Protection Regulation (GDPR), which ensures that patient data is safeguarded with strict privacy protocols. The EU also has a continuous monitoring requirement.

In the United States of America, the Food and Drug Administration (FDA) plays a central role in regulating AI in healthcare. The FDA treats AI as Software as a Medical Device (SaMD) and applies stringent safety, performance, and accountability requirements similar to those governing traditional medical devices. The FDA has a Digital Health Innovation Action Plan that focuses on balancing rapid innovation in AI with robust oversight to ensure patient safety.

AI systems must pass rigorous premarket reviews before deployment, and once deployed, they are subject to ongoing monitoring. The FDA has also developed a total product lifecycle approach, ensuring that AI systems continue to meet safety standards as they evolve and interact with new patient data. This regulatory framework supports innovation while maintaining a strict focus on transparency, fairness, and patient protection.

The United Kingdom's National Health Service (NHS) has established its own framework for AI safety through the NHS AI Lab. This initiative supports the ethical integration of AI technologies into the healthcare system. The NHS provides a mechanism to test AI before being implemented. It has an AI in Health and Care Award to facilitate funding and support for AI projects that align with clinical needs. It has also developed an AI Ethics Initiative within the NHS. This framework also prioritises patient engagement, ensuring that AI systems are developed with input from those who will be most affected, such as patients and clinicians. The UK has retained (GDPR) requirements.

Finland is also actively engaged in creating a comprehensive AI strategy, with a specific focus on healthcare. The country has adopted a Human-Centric AI Strategy, which emphasizes fairness, transparency, and the ethical deployment of AI technologies. Finland's approach is deeply collaborative, involving healthcare providers, AI developers, and policymakers to ensure that AI systems are designed to serve patients in an inclusive and fair manner. Finland also focuses on ensuring that AI systems are explainable and that healthcare professionals receive adequate training on how to integrate AI into clinical workflows.

Canada has developed the Pan-Canadian AI Strategy, which addresses the ethical deployment of AI across sectors, including healthcare. This Canadian Institute for Advanced Research (CIFAR) developed the strategy which mentions fairness, inclusivity, and the ethical use of AI. In healthcare, the strategy stresses the importance of ensuring that AI technologies do not exacerbate existing inequalities, particularly for underserved or vulnerable populations. Additionally, the framework promotes transparency, bias mitigation,

and continuous monitoring to ensure that AI systems in healthcare align with Canadian values of equity and fairness.

Health and AI Consultation Findings

In late 2023, SPG held consultations with migrant women and youth across Melbourne, Sydney, and Brisbane to understand their views on the integration of AI into healthcare. These sessions, comprising 7 to 12 participants ranging in age from 19 to 66, revealed a wide spectrum of awareness regarding AI technology and its potential use in healthcare.

While some participants had little to no understanding of AI, others had a general grasp of its applications but expressed concerns about how it would function in practice. A recurring fear was that AI might replace their doctor, leading to an impersonal healthcare experience. Many participants voiced apprehension about losing the human connection with their health care provider, a particularly important aspect for those already navigating cultural and language barriers within the Australian healthcare system.

One of the most frequently raised concerns in all three consultations was whether AI systems would be able to communicate in their native languages and understand cultural nuances essential for their healthcare. Older participants, in particular, worried that language limitations might prevent them from receiving accurate care from AI-driven systems. Additionally, several participants questioned whether AI would be tailored to address healthcare issues specific to women, fearing that male- dominated medical data might lead to the neglect of conditions that disproportionately affect women or minority ethnic groups. This view was particularly prominent among young women from migrant backgrounds.

Overall, women expressed concerns that AI systems, if not inclusive, might exacerbate health inequities instead of alleviating them. Another significant concern that surfaced was the issue of data privacy. Many participants feared that their personal medical information might not be adequately protected by AI systems, potentially leading to data breaches or misuse of sensitive health information.

These interviews revealed a persistent sense of fear among participants, especially regarding AI's ability to handle cultural contexts and personal healthcare preferences. However, a few participants expressed cautious optimism. They highlighted the potential of AI to enhance communication and simplify access to healthcare services, particularly by removing existing language barriers.

Following these consultations, SPG conducted deep dive interviews with some of the participants. Further interviews were conducted with bicultural support workers, and bicultural clinicians to gain deeper insights. Bicultural support workers and clinicians echoed the views from the consultations, noting that while AI could improve efficiency, it is critical for AI systems to be designed with cultural sensitivity and input from diverse communities to ensure equitable healthcare outcomes. Additionally, they reiterated the importance of robust data privacy measures, suggesting that failure to secure patient data could further erode confidence in AI technologies among migrant communities and could undermine trust between patients and their doctors.

Designing AI Solutions for Inclusivity and Equity

The ultimate goal of SPG's work is to ensure that AI systems are designed with inclusivity and equity as core principles. This involves moving beyond simply avoiding harm and actively working to reduce disparities and increase access. To achieve this, SPG advocates for the following design principles:

Community-Centred Design*:* AI systems that are procured by the Government should be developed with input from the communities they are meant to serve. This includes involving community representatives in the design, testing, and implementation of AI systems. By doing so, AI developers can ensure that the tools they create are relevant and responsive to the specific needs of vulnerable populations. Additionally, involving the community in the design process helps to build trust in AI technologies, ensuring that patients feel comfortable using these tools.

Linguistic and Cultural Flexibility*:* AI systems must be designed to accommodate linguistic diversity and cultural differences. This includes providing support for multiple languages, as well as ensuring that the AI system can interpret cultural variations in health beliefs, practices, and symptom descriptions. For example, an AI system designed for mental health support should be able to recognise culturally specific expressions of mental distress and provide care that is culturally appropriate.

Bias Auditing and Ethical Oversight*:* AI systems in healthcare must be regularly audited to ensure that they are not reinforcing existing biases. This includes conducting bias audits to assess how the system performs across different demographic groups and making adjustments to the algorithm if disparities are identified. Ethical oversight committees,

including representatives from vulnerable communities, should be established to ensure that AI systems are aligned with ethical standards and principles of equity.

The considered design of principles that ensure AI-driven healthcare solutions contribute to a more equitable and inclusive healthcare system requires considerable forethought. Through comprehensive consultations, partnerships with health peak organisations, and a focus on addressing intersectional barriers, SPG proposes to work to reduce healthcare disparities and ensuring that all individuals, regardless of their background, have access to high-quality, culturally competent care, protecting medical professionals and patients.

Addressing intersectional barriers in healthcare requires a comprehensive, multi- sectoral that must be led by both clinical experts and community. Further community consultation, partnership formation, and the design of inclusive AI-driven healthcare solutions is the first step. SPG's demonstrated experience in identifying and addressing these barriers positions it as a key advocate for vulnerable populations in the development and implementation of AI technologies in healthcare. By ensuring that AI systems are designed with cultural and linguistic competency, inclusivity, and equity at their core, SPG contributes to the creation of a healthcare system that serves all individuals, particularly those from marginalised and underserved communities.

Comprehensive Review of the 10 Voluntary AI Safety Guardrails, Gender, Intersectionality, and Recommendations

Australia's Voluntary AI Safety Standards outline 10 guardrails designed to promote safe, fair, and accountable AI use. While these standards are voluntary, they may eventually become mandatory for high-risk AI applications. While this is outside the consultation scope, SPG have conducted a detailed analysis of the benefits, gaps, and recommendations for each guardrail, with a focus on how these standards should apply to medium- and lower-risk AI systems, especially in protecting marginalised people and communities and other intersectional groups.

Guardrail 1:

Establish, implement, and publish an accountability process, including governance, internal capability, and a strategy for regulatory compliance

Benefits: Accountability ensures that organisations implement clear structures for managing the development, deployment, and use of AI systems. This guardrail calls for organisations to assign ownership of AI systems, create an AI strategy, and ensure compliance with regulatory standards. Having a formalised accountability process promotes transparency and helps mitigate potential risks.

Gaps: While it ensures general oversight, the guardrail does not mandate intersectional impact assessments within its accountability framework. This can lead to women, such as migrant women, being overlooked in decisions related to AI development and governance. Without a focus on intersectional fairness, AI systems might disproportionately affect vulnerable groups, particularly in employment or public services.

International Comparison: The EU AI Act requires mandatory accountability frameworks for high-risk AI systems, including bias audits. This ensures that bias detection is built into governance processes, promoting fairness and inclusivity in AI development and deployment.

Recommendations:

1. Mandatory for medium-risk systems: Accountability should be required in medium-risk AI systems (such as recruitment, education, or telecommunications) where bias can have significant social and economic consequences.
2. Intersectional impact assessments should be incorporated within accountability processes to ensure that marginalised groups, including migrant women and women of colour, are not disproportionately harmed.

Guardrail 2:

Establish and implement a risk management process to identify and mitigate risks

Benefits: This guardrail ensures that organisations identify and mitigate the risks associated with AI systems before and during deployment. Risk management processes are vital in high-risk settings like healthcare or employment, where AI systems could significantly affect people's lives.

Gaps: Without intersectional risk assessments, organisations may overlook how certain groups—such as migrant women, women of colour, or LGBTIQ+ individuals—are uniquely

vulnerable to AI risks. Medium-risk AI systems (such as those used in public administration or recruitment) also need thorough risk assessments, as these settings can still produce cumulative harms for marginalised groups.

International Comparison: Canada's Pan-Canadian AI Strategy requires risk assessments that consider the impact on marginalised populations, ensuring that AI systems do not disproportionately affect vulnerable communities.

Recommendations:

3. Mandatory for medium-risk systems: Risk management should be required in medium-risk applications, such as education, recruitment, and public services, where AI systems may disproportionately harm marginalised people and communities.
4. Require intersectional risk assessments to ensure that organisations actively assess how AI systems affect women of colour, migrant women, and LGBTIQ+ individuals.

Guardrail 3:

Protect AI systems and implement data governance measures to manage data quality and provenance

Benefits: Ensuring data governance is essential to protect the quality, provenance, and security of data used in AI systems. Good data governance practices help prevent biased AI systems from being trained on incomplete or inaccurate data, reducing the risk of discriminatory outcomes.

Gaps: The current guardrail does not mandate the use of diverse datasets to ensure that AI systems represent the full breadth of society. This omission is particularly problematic for medium-risk applications, where AI systems may perpetuate bias in sectors like education, customer service, or recruitment, if trained on unrepresentative data.

International Comparison: The OECD AI Principles highlight the need for inclusive data governance across all AI systems to prevent the development of biased or discriminatory AI models.

Recommendations:

5. Mandatory for all AI systems: Data governance should be required in all AI systems, regardless of risk level, as biased data can lead to significant harm even in low-risk applications.
6. Ensure intersectional data audits to verify that AI systems are trained on inclusive datasets that reflect the diversity of society, including marginalised people and communities.

Guardrail 4:

Test AI models and systems to evaluate model performance and monitor the system once deployed

Benefits: Regular testing of AI models ensures they perform as intended and do not produce harmful biases. Monitoring AI systems after deployment allows organisations to detect and address any unintended consequences over time.

Gaps: This guardrail primarily focuses on high-risk AI systems, but medium- and lower-risk systems should also undergo testing to ensure they do not create cumulative harm. For instance, automated grading systems in education or AI-driven hiring tools may unfairly disadvantage migrant women or women of colour, yet these systems may not undergo rigorous bias testing if classified as lower-risk.

International Comparison: The EU AI Act requires thorough pre-market testing and post-market monitoring to detect biases in AI systems before and after deployment, ensuring that all groups are treated fairly.

Recommendations:

7. Mandatory for medium-risk systems: Testing and monitoring should be required for medium-risk AI systems, such as hiring platforms, educational software, and public services.
8. Conduct intersectional bias testing to ensure that AI systems perform equitably across all demographic groups, including marginalised women and LGBTIQ+ individuals.

Guardrail 5:

Enable human control or intervention in an AI system to achieve meaningful human oversight across the lifecycle

Benefits: Human oversight ensures that AI systems are correctable and can be adjusted if they produce harmful or unintended outcomes. This is crucial for preventing AI systems from making discriminatory decisions without human intervention.

Gaps: Although human oversight is critical in high-risk applications, medium- and lower-risk AI systems also need mechanisms for human intervention. For instance, AI systems used in customer service or education may produce biased outcomes, but without human oversight, these issues may go unnoticed. Additionally, there is no requirement for bias training for those overseeing AI systems, leaving potential intersectional harms undetected.

International Comparison: In the United States Food and Drug Administration (US FDA) guidelines for AI in healthcare, human oversight is coupled with training on bias detection, ensuring that those overseeing AI systems are equipped to recognise and correct discriminatory outcomes.

Recommendations:

9. Mandatory for medium-risk systems: Human oversight should be required in medium-risk AI systems, such as educational tools, hiring platforms, and public administration.
10. Require intersectional bias training for those overseeing AI systems to ensure that marginalised and other vulnerable groups are not disproportionately impacted by AI decisions.

Guardrail 6:

Inform end-users regarding AI-enabled decisions, interactions with AI, and AI-generated content

Benefits: Transparency is critical for ensuring that end-users understand when AI is involved in decision-making processes. This guardrail promotes user trust by requiring organisations to inform individuals about the role AI plays in generating decisions or content.

Gaps: Medium- and lower-risk AI systems also require transparency, particularly for marginalised people and communities who may be unaware that AI is influencing their employment applications, loan approvals, or service interactions. If end-users are not informed, they cannot challenge biased outcomes or seek clarification.

International Comparison: The OECD AI Principles emphasise accessible communication, ensuring that AI-related information is presented in a way that is understandable to non-native speakers or individuals with limited technical knowledge.

Recommendations:

11. Mandatory for medium-risk systems: Transparency requirements should be extended to medium-risk AI systems, including recruitment tools, education platforms, and public services.
12. Ensure that AI-related information is provided in multiple languages and accessible formats to accommodate the needs of marginalised people and communities.

Guardrail 7:

Establish processes for people impacted by AI systems to challenge use or outcomes

Benefits: This guardrail provides individuals with a formal mechanism to challenge AI decisions, ensuring that people affected by AI systems can seek recourse if they feel an outcome is unfair or biased. This process is essential for high-risk AI systems, such as those used in employment or law enforcement, where the consequences of AI decisions can be severe.

Gaps: Even in medium-risk applications—such as education, financial services, and public administration—individuals may face biased AI decisions. Without the ability to challenge outcomes, marginalised people and communities may be disproportionately affected by automated decisions that are influenced by unconscious biases within the AI system. Additionally, the challenge process must be accessible, particularly for migrant women or those with limited English proficiency, to ensure that they can effectively engage with the system.

International Comparison: The EU AI Act guarantees the right to an explanation and provides mechanisms for challenging AI decisions, ensuring that individuals are given clear information and pathways to contest biased outcomes.

Recommendations:

13. Mandatory for medium-risk systems: Processes for challenging AI decisions should be required for medium-risk systems, including AI-driven loan approvals, recruitment platforms, and educational tools.
14. Ensure that challenge processes are accessible in multiple languages and culturally competent, allowing marginalised people and communities and those with limited technical knowledge to effectively contest AI outcomes.

Guardrail 8:

Be transparent with other organisations across the AI supply chain about data, models, and systems to help them effectively address risks

Benefits: Transparency within the AI supply chain is essential for ensuring that data, models, and systems used by AI are shared across organisations to mitigate risks. When organisations understand how an AI system was built and the data it was trained on, they can better address potential biases and harmful outcomes.

Gaps: While transparency is crucial for high-risk applications, it should also apply to medium- and lower-risk AI systems. For instance, AI systems used in retail, customer service, or education may rely on third-party datasets that contain hidden biases against marginalised people and communities. Without adequate transparency across the supply chain, organisations may unknowingly adopt AI systems that perpetuate discrimination or exclude vulnerable groups.

International Comparison: The OECD AI Principles highlight the need for supply chain transparency to prevent unintended harms and promote ethical use of data throughout the AI lifecycle.

Recommendations:

15. Mandatory for medium-risk systems: Supply chain transparency should be required for medium-risk applications, particularly in education, recruitment, and consumer services, where third-party AI tools are often used.
16. Ensure that data-sharing practices protect the privacy of marginalised and vulnerable people and communities who may face additional risks if their data is shared inappropriately across supply chains.

Guardrail 9:

Keep and maintain records to allow third parties to assess compliance with guardrails

Benefits: Record-keeping allows organisations to demonstrate their compliance with AI safety guardrails. By maintaining detailed records of their AI systems, data, and decision-making processes, organisations can provide audit trails for regulators or stakeholders to assess whether the AI is operating in a fair and transparent manner.

Gaps: Record-keeping is important not only for high-risk AI systems but also for medium- and lower-risk systems. For example, AI-driven recruitment systems may not be considered high-risk, but if they disproportionately impact migrant women or women of colour, records will be necessary to assess whether bias was introduced during the system's development or deployment.

International Comparison: The EU AI Act mandates comprehensive record-keeping for high-risk AI systems, ensuring that organisations maintain detailed records for accountability and transparency. The OECD AI Principles also underscore the importance of record-keeping for demonstrating compliance with ethical standards.

Recommendations:

17. Mandatory for medium-risk systems: Record-keeping should be required for medium-risk AI systems that influence recruitment, education, and public services, as these areas can have long-term effects on marginalised communities.
18. Ensure that records include bias audits and intersectional impact assessments, allowing third parties to assess whether AI systems disproportionately harm marginalised women or other vulnerable groups.

Guardrail 10:

Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion, and fairness

Benefits: Engaging with stakeholders helps organisations better understand the potential impacts of AI systems on different communities. This guardrail encourages organisations to prioritise safety, diversity, inclusion, and fairness by actively involving stakeholders in AI governance processes.

Gaps: While stakeholder engagement is essential, it does not mandate engagement with marginalised groups, such as migrant women, First Nations people, or LGBTIQ+ individuals. Without proactive efforts to include these groups, organisations risk deploying AI systems that fail to reflect the needs and concerns of vulnerable populations. In medium-risk applications, such as education or public services, engaging with diverse stakeholders is crucial to ensure AI systems are inclusive.

International Comparison: Canada's Pan-Canadian AI Strategy mandates community engagement to ensure that underrepresented groups are involved in the design and deployment of AI systems.

Recommendations:

19. Mandatory for all AI systems: Stakeholder engagement should be required for all AI systems, including medium- and lower-risk applications, to ensure that the needs of marginalised groups are considered in AI development.
20. Actively involve representatives from intersectional groups—such as women of colour, migrant women, and First Nations communities—in stakeholder engagement processes to ensure that their perspectives inform AI governance.

Conclusion

Australia's current approach to AI safety, framed by voluntary guardrails, reflects an understandable ambition to harness the potential of AI-driven innovations in healthcare, employment, and other critical sectors. However, by prioritising the economic benefits of AI—such as productivity gains and contributions to GDP growth—the framework risks overlooking the profound social and economic disruptions that AI is likely to cause. This includes the anticipated structural changes in Australia's consumption-led economy, particularly the displacement of large sections of the middle-income workforce, raise significant concerns, especially for women and marginalised communities.

While the voluntary guardrails provide a foundation for AI governance, the decision to limit regulation to high-risk settings and the adoption of a one-size-fits-all approach create a relatively weak regulatory framework for AI safety. This narrow focus leaves critical gaps in addressing the broader risks that AI poses, particularly in sectors that may initially appear low-risk but can become high-risk over time. Furthermore, limiting regulation to high-risk settings overlooks the compounded harms faced by for example, marginalised women, migrant communities, and other vulnerable groups across a variety of contexts.

As AI technologies become more prevalent, the need for comprehensive regulatory frameworks that enforce safety, accountability, and fairness across all sectors is critical. The current voluntary guardrails do not sufficiently address the potential for AI to exacerbate intersectional inequalities, particularly among migrants, people of colour, LGBTIQ+ individuals, and First Nations. The lack of enforceability further weakens the framework's ability to ensure that AI systems are safe, inclusive, and non-discriminatory from the development stage through to deployment.

A robust regulatory ecosystem that includes auditing at the development stage, training for developers and users, and continuous monitoring of AI systems is essential to mitigate the risks posed by AI, particularly in high-risk sectors. Moreover, the issue of skill attrition and the potential displacement of workers due to automation—especially in fields like healthcare, engineering, and public services—must be addressed through comprehensive upskilling and reskilling initiatives.

In conclusion, while AI holds the promise of substantial economic gains and technological advancement, Australia's AI governance framework must evolve to balance these opportunities with a clear-eyed focus on the societal impacts. A more enforceable, intersectionally aware regulatory regime—one that prioritises both innovation and the protection of women and vulnerable communities—is needed to ensure that AI serves the needs of all Australians equitably. Only by addressing these challenges can Australia position itself as a leader in ethical AI governance, capable of maximising the benefits of AI while safeguarding against its potential harms.

Australia's ten voluntary guardrails provide a strong foundation for responsible AI governance. However, there are significant gaps when it comes to addressing the concerns of marginalised and other intersectional groups. Many of these guardrails, while currently focused on high-risk applications, should also be made mandatory for medium- and lower-risk AI systems, where the potential for cumulative harms is still significant. By adopting a more intersectional approach and extending these protections, Australia can ensure that its AI governance framework promotes fairness, inclusivity, and equity for all.

References:

1. Australian Commission on Safety and Quality in Health Care (ACSQHC). AI in Clinical Practice: Safety and Quality Guidelines, 2022. Available at: [URL]
2. Australian Department of Health and Aged Care. AI in Healthcare Initiatives, 2022. Available at: [URL]/
3. Australian Digital Health Agency (ADHA). Strategic Plan for AI Integration in Healthcare: Ethics and Governance, 2020. Available at: [URL]
4. Australian Government Department of Health and Aged Care. (2022). Artificial Intelligence in Healthcare: A National Strategy. Canberra: Department of Health and Aged Care.
5. Australian Government Department of Industry, Science and Resources. (2023). Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings. Canberra: Department of Industry, Science and Resources.
6. Australian Government Department of Industry, Science and Resources. (2024). Voluntary AI Safety Standard. Canberra: Department of Industry, Science and Resources.
7. Australian Human Rights Commission. (2021). Human Rights and Technology Final Report. Sydney: AHRC.
8. Australian Human Rights Commission (AHRC). Human Rights and AI: The Intersection with Healthcare, 2020. Available at: [URL]/
9. Australian Medical Association (AMA). AI in Healthcare: Clinical Safety and Guidelines, 2020. Available at: [URL]
10. Australian Alliance for Artificial Intelligence in Healthcare (AAAiH). AI Roadmap for Safe and Ethical AI Implementation in Health Systems, 2021. Available at: [URL]
11. Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency. New York: ACM.
12. Binns, R. Fairness in Machine Learning: Lessons from Political Philosophy, Proceedings of the Conference on Fairness, Accountability, and Transparency, 2018.
13. Brundage, M., & Bryson, J. J. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Journal of Cybersecurity, 4(1), 1-13.
14. Canada's Digital Technology Supercluster. (2017). Pan-Canadian Artificial Intelligence Strategy. Ottawa: Digital Technology Supercluster.
15. Coiera, E. & Verspoor, K. AI in Healthcare: Gaps in Australia's Preparedness and Recommendations, Medical Journal of Australia, 2023.

16. Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven, CT: Yale University Press.
17. Data61, CSIRO. Guidelines for Responsible AI in Health: Operationalising Ethics, 2019. Available at: [URL]
18. European Commission. (2021). Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Brussels: European Commission.
19. European Commission. Ethics Guidelines for Trustworthy AI, 2019. Available at: [URL]
20. European Union Agency for Fundamental Rights (FRA). (2020). Getting the Future Right – Artificial Intelligence and Fundamental Rights. Luxembourg: Publications Office of the European Union.
21. Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press.
22. Ethics Centre Australia. Frameworks for Ethical AI in Healthcare: Challenges and Solutions, 2020. Available at: [URL]/
23. Fenech, M., Strukelj, N., & Buston, O. AI in Healthcare: The Ethical, Social, and Governance Challenges, Cambridge University Press, 2020.
24. Future of Life Institute. (2017). Asilomar AI Principles. Future of Life Institute.
25. Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. IEEE Internet Computing, 21(6), 58-62.
26. Gasser, U., & O'Reilly, P. (2020). AI and Human Rights: A Human Rights-Centered Approach to Artificial Intelligence Governance. Harvard University: Berkman Klein Center for Internet & Society.
27. Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1(9), 389-399.
28. Mittelstadt, B. (2019). Principles Alone Cannot Guarantee Ethical AI. Nature Machine Intelligence, 1(11), 501-507.
29. Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. The Ethics of Algorithms: Mapping the Debate, Big Data & Society, 2016.
30. Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press.
31. Obermeyer, Z., & Mullainathan, S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, Science, 2019.
32. OECD. (2019). Recommendation of the Council on Artificial Intelligence. Paris: OECD Publishing.
33. Royal Australian and New Zealand College of Psychiatrists (RANZCP). AI in Mental Health: Guidelines for Ethical Implementation, 2020. Available at: [URL]/

34. Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Upper Saddle River, NJ: Pearson.
35. The Lancet Digital Health. Artificial Intelligence in Healthcare: A Systematic Review of Ethical Issues, 2021. Available at: [URL]
36. Taddeo, M., & Floridi, L. (2018). How AI Can Be a Force for Good. Science, 361(6404), 751-752.
37. The National Health Service (NHS) UK. AI in Health and Care: Implementation and Ethics, 2021. Available at: [URL]
38. The Therapeutic Goods Administration (TGA). Regulation of Software-based Medical Devices including AI, 2021. Available at: [URL]
39. Topol, E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Basic Books, 2019.
40. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO Publishing.
41. UNESCO. AI and Human Rights in Healthcare: Global Ethical Recommendations, 2020. Available at: [URL]
42. United States Food and Drug Administration (FDA). (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. Washington, DC: FDA.
43. United States Food and Drug Administration (FDA). Artificial Intelligence and Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Action Plan, 2021. Available at: [URL]/
44. World Health Organization. (2021). Ethics and Governance of Artificial Intelligence for Health. Geneva: WHO.
45. World Health Organization (WHO). Ethical Considerations for AI in Healthcare: Global Guidelines, 2021. Available at: [URL]
46. West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race, and Power in AI. New York: AI Now Institute.