Actuaries Institute.

2 October 2024

Department of Industry, Science and Resources
Industry House
10 Binara Street
Canberra ACT 2601

Email: aiconsultation@industry.gov.au

Dear Sir/Madam,

Department of Industry, Science and Resources Consultation on mandatory guardrails for safe and responsible AI

The Actuaries Institute ('the Institute') welcomes the opportunity to respond to the proposals paper for introducing mandatory guardrails for AI in high-risk settings.

The Institute is the peak professional body for actuaries in Australia. Our members work in a wide range of fields including insurance, superannuation and retirement incomes, banking, enterprise risk management, data science and AI, climate change impacts and government services. The Institute has a longstanding commitment to contribute to public policy discussions where our members have relevant expertise. The comments made in this submission are guided by the Institute's 'Public Policy Principles' that any policy measures or changes should promote public wellbeing, consider potential impacts on equity, be evidenced-based and support effectively regulated systems.

Acknowledgment of progress in AI regulation proposals

The Institute welcomes and acknowledges the progress made by the Australian Government in refining its approach to AI regulation since the 2023 Safe and Responsible AI discussion paper. The current proposal demonstrates a more nuanced understanding of the complexities involved in regulating AI systems, reflecting some of the concerns raised by stakeholders, including the Institute, in previous consultations.

One notable area of improvement is the adoption of a more widely accepted definition of AI systems. The current proposal uses the well accepted OECD definition without alteration (p. 8), addressing our previous concern about overly narrow or confusing bespoke definitions. This change provides a clearer foundation for the regulatory framework, reducing potential ambiguity in its application, and helps with international alignment.

Another significant advancement is the more detailed articulation of risk dimensions for AI systems (pp. 19-29). The current proposal outlines specific principles for determining high-risk AI, including impacts on human rights, health and safety, and broader societal effects. This represents a substantial improvement over the previous paper's more general approach, aligning more closely with our recommendation for a well-defined taxonomy of risks.

The introduction of guardrails for high-risk AI systems (pp. 35-42) is a third area of progress. These guardrails provide a clearer framework for developers and deployers to follow than the previous proposal, addressing our previous call for more specific risk management options. While we believe further refinement of these is necessary, this represents a positive step towards practical, implementable regulation.

Actuaries Institute
Level 2, 50 Carrington Street, Sydney NSW 2000
P +61 (0) 2 9239 6100 | actuaries.asn.au
ABN 69 000 423 656

Further refinement needed to get to effective AI regulation

Despite these improvements, there remain areas where further progress is needed.

- The current risk classification system still lacks the specificity we believe is necessary for effective regulation.
- The proposed guardrails, while more specific than interventions proposed in last year's consultation, may not sufficiently address the full range of AI-related risks that need to be captured – both those we have identified in the past, and those that have been detailed in this proposals paper.
- The application of guardrails may create excessive costs and / or externalities, if implemented in a broad manner.

We restate our general recommendations for AI regulation made in response to the 2023 Safe and Responsible AI discussion paper, which we believe still apply and are relevant to this consultation. In our detailed comments in the attachment, we examine the current proposals paper considering those recommendations, identifying further steps which we feel are still required, particularly if certain regulatory choices are made. We do this both in general terms and by way of stylised examples.

In our response to the 2023 Safe and Responsible AI discussion paper, we made four general recommendations which are still are applicable:

1. *Regulation should primarily be outcomes focused, rather than technology focused, to help ensure it can be enduring/long lasting.*

   We repeat this recommendation. As outlined in detail in our previous response, AI specific regulation carries with it a range of structural challenges which are avoidable if other approaches are chosen.

   Our concern here is more relevant if regulatory option three is chosen, as then the definition of AI has direct application. It may be less of a concern under options one and (perhaps) two, but this would depend on how those options are pursued.

   However, we welcome the current consultation paper's use of the well-accepted AI System definition from the OECD, unaltered. The use of a standard definition will reduce some (but not all) of the issues which will be faced if this is applied directly.

   In the case of General Purpose AI (GPAI), there may be a case for specific regulation. However, we consider that the definition proposed in the discussion paper is too broad to make that feasible. If GPAI is defined more narrowly, and if the risks of these systems are clearly identified, there may be a case to be made for GPAI-specific regulation that appropriately targets the unique risks these systems may pose.

2. *Risk-based approaches to AI regulation should:*
   - *be based on a well-defined taxonomy of risks that AI systems may introduce or exacerbate;*
   - *incorporate a well-defined menu of risk management options that could be imposed by regulation;*
   - *ensure the costs of risk-based regulatory interventions are justified by the risk reduction created, without obvious gaps or overreach; and*
   - *carefully target risk management interventions to the risks identified for each situation considered, rather than bluntly applying the same interventions across a broad, vaguely defined risk category.*

The current consultation demonstrates improved maturity compared to a year ago. Dimensions of 'risk' have been articulated which can now be examined, and interventions ('guardrails') are outlined in a little more depth. However:

- The 'risk' framework will cause challenges if applied directly, which we discuss in section 2 of the Attachment. Again, this poses more challenges under option three than options one and two.
- The structure and, in some places, content of the guardrails is still concerning. We feel there is much more to be done if we are to avoid the gaps or overreach which we highlighted in 2023 as a limitation of this sort of model. We illustrate with examples some of the issues that the current approach will create, in section 3 of the Attachment, and suggest ways these challenges might be avoided.

3. *Producing guidance on existing regulation should be prioritised over creating new regulation, in situations where such regulation already exists.*

4. *A centralised expert body should be created and appropriately funded, to provide assistance to primary regulators in considering AI governance, regulation and guidance.*

Our previous recommendations (above) (3) and (4) remain. Irrespective of which regulatory option is selected following this consultation, existing regulation already requires clarifying guidance. We are concerned that – particularly if regulatory option three for implementing the guardrails is selected – this will not be prioritised. This would be an error. The creation of *ex ante* guardrails does not help to clarify the operation of existing regulation, particularly regulation which operates *ex post*. A centralised expert body to help produce such guidance is even more relevant today than a year ago.

Ongoing assistance

The Institute remains committed to working collaboratively with Government to further refine and improve the AI regulatory framework. We offer our expertise in risk assessment and management, particularly in areas such as insurance and financial services, to help develop more targeted and effective regulatory approaches. Our members stand ready to participate in working groups, provide detailed feedback on technical aspects of the proposals, or contribute to pilot programs testing the proposed regulatory mechanisms.

If you would like to discuss any aspect of this submission please contact the Institute via (02) 9239 6100 or public_policy@actuaries.asn.au.

Yours sincerely

(Signed) Elayne Grace
CEO

Attachment: Detailed comments

1. Regulating 'AI'

As outlined in detail in our response to the 2023 consultation, AI specific regulation begs obvious questions – why is intervention required for AI systems, but not for equivalent non-AI systems? What challenges does that inconsistency create? Certainly, there are high stakes non-AI decisions being made every day which can cause harm. This includes simple instances of automation which would not be considered 'AI' (e.g. robodebt), and high stakes human decisions throughout the economy. As we have said previously, separation of regulatory requirements by way of an AI definition will also cause challenges for hybrid systems, particularly where AI plays only a very small part in the overall decision-making process.

The proposals paper allows a more thorough illustration of this point. We can examine the proposed guardrails, and consider whether they could, and perhaps should, be applied in general rather than just to AI Systems.

Considering the guardrails in turn:

Guardrail	**Consideration of broader application**
1.Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance	If a decision meets the threshold of being 'high risk' but is made by non-AI means, is it less important to have clear accountability? We suggest the need for accountability is driven by the impact of the decision, not the mechanism by which the decision is made. This Guardrail could apply more broadly.
2. Establish and implement a risk management process to identify and mitigate risks	Again, this appears to be a sensible general requirement imposed on an organisation making high stakes decisions that ought not to be restricted to AI – non-AI decisions clearly also require risk management.
3. Protect AI systems, and implement data governance measures to manage data quality and provenance	Data governance mechanisms are perhaps more important for AI Systems than others, but they are still necessary in many non-AI situations. For example, a credit application process involving human assessment and decisions must still ensure the information provided within the application is suitably managed and secured.
4. Test AI models and systems to evaluate model performance and monitor the system once deployed	Non-AI decisions also require evaluation. The human equivalent of this guardrail is to test the standard procedures, rules or authorities given to human decision makers and to monitor the quality of decisions in aggregate and/or through a quality assurance process. While somewhat different in style, the spirit of the intervention is similar and surely equally important for non-AI decisions as for AI ones.
5. Enable human control or intervention in an AI system to achieve meaningful human oversight	Though again it is different in style, a process for independent checking or quality assurance of high stakes human decisions is already considered good practice in many sectors. For example, high risk credit applications may be subject to review by a manager, after initial assessment by a junior team member.

6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content	This guardrail is intended to provide transparency about how a decision is made or how an output is generated. Again, if the application is a high-risk one, we see no reason to restrict this only to AI – if people have the right to know how a decision is made or an output generated, this should be a general one.
7. Establish processes for people impacted by AI systems to challenge use or outcomes	This is already common for many high stakes decisions, with internal and external dispute systems in place in many sectors. Again, restricting this only to AI (and presumably requiring incremental action only where those mechanisms are not already in place) appears too limited. All high stakes decisions warranting dispute mechanisms should have them, irrespective of whether those decisions are made via AI or otherwise. Why would we create such mechanisms only for AI-based decisions?
8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks	This guardrail is more particular to the AI supply chain and perhaps in this case there is no human analogue.
9. Keep and maintain records to allow third parties to assess compliance with guardrails	Record keeping is also good practice in high stakes non-AI decisions, particularly in situations where decisions are to be reviewed or contested.
10. Undertake conformity assessments to demonstrate and certify compliance with the guardrail	Noting the comments above, conformity assessments for high stakes non-AI processes could equally be imposed, to ensure those processes are working as intended

Based on the above comments, if these guardrails are to be imposed across the economy as *ex ante* protections against harms in high stakes situations, then with the exception of Guardrail (8), we can see no reason why the guardrails should be limited to only AI-generated decisions (i.e., the guardrails should apply to high stakes decisions in general, whether AI, hybrid, human generated, or anything else). To not do this risks confusion, particularly where a decision process involves both AI and non-AI components. It will also discourage innovation, as AI-Systems will have more onerous requirements placed over them than non-AI ones, even if the decisions being made are exactly the same.

2. Defining High-Risk AI

The current consultation demonstrates improved maturity in defining the area of application compared to a year ago. However, we still believe there is significant room for improvement.

If applied directly, the definition of “high-risk” articulated will lead to:

- gaps, as it will fail to capture things which ought to be captured;
- waste, as it will capture things which ought not to be captured; and
- confusion and inconsistent application, as in several places it is described in vague or confusing terms which cannot be confidently applied in practice.

These problems are more easily avoidable if regulatory option one is pursued, or under certain structures of option two, since this may allow suitable interpretation as primary legislation is drafted, rather than being applied directly. Under option three, these issues may be severe though there may still be mitigants available, which we discuss further below.

2.1 Potential Gaps

2.1.1 Economic Impact

AI systems can have significant financial or economic impacts on people. It is unclear if this is intended to be captured under the risk/impact dimensions proposed.

For example, proposed principle (c)[1] discusses 'legal effects' primarily in the context of restricting access to services, but (for example) an AI system used to set prices for flood insurance would not necessarily result in a refusal to sell a policy but might result in a price that is unaffordable to some. This sort of effect might be captured under principles (d) or (e), but again this is not entirely clear.

We suggest that Government either include economic impacts as a separate dimension to consider, or else clarify if and how economic impacts are intended to be captured within the proposed framework.

In doing so, we suggest a clear threshold be identified, as while some AI systems may have a significant economic impact, others might be modest and perhaps not worthy of the same attention.

Similarly, AI systems may be used to personalise services offered, which may not meet the 'legal effects' definition within the framework but could still have a significant impact on people. For example, an insurance contract may have special terms imposed based on an AI system output. The framework should contemplate these sorts of situations and clarify if and how they are captured.

If a new dimension is to be created, we suggest the following form of wording might be suitable to consider:

g) Economic and financial risks, such as access to finance or the costs or nature of essential services made available.

2.1.2 Cybersecurity

Some AI systems may introduce novel cybersecurity threats. This is an emerging area of consideration particularly regarding Large Language Model based applications, which have been shown to be vulnerable to various novel attacks. This does not appear to be captured by the dimensions shown which are primarily centred on the direct impact of an AI system on individuals, groups or society.

The introduction of novel cybersecurity threats is a less direct, but still relevant, area of risk to consider.

Without this inclusion, AI systems with significant cybersecurity threats could be described as 'low risk'.

We suggest that a new dimension could be created to capture this risk, and that international work on this topic such as the *OWASP Top 10 for Large Language Model Applications*[2] should be used as a reference point for both the threats and potential mitigations.

2.1.3 Privacy and Data Protection

Similarly to cybersecurity, the current risk framework does not adequately address the specific privacy and data protection concerns that AI systems can introduce or exacerbate. While some aspects of privacy might be covered under 'legal effects', the unique challenges posed by AI in this domain warrant separate consideration.

AI systems often process vast amounts of information, sometimes personal or sensitive information, potentially exposing individuals to increased risks of privacy breaches, identity theft and unauthorised data use. These risks go beyond traditional data protection concerns due to the ability of some AI

[1] Proposal Paper, page 19

[2] https://genai.owasp.org/llm-top-10/

systems to infer sensitive information from seemingly innocuous data, create detailed profiles, and potentially re-identify data that was thought to be anonymised.

For example:

- An AI system used for personalised marketing could inadvertently reveal sensitive personal information by inferring health conditions or financial status from browsing patterns or purchase history.
- Large Language Models trained on public data might reproduce or allow extraction of personal information that was incidentally included in their training data.
- AI-powered facial recognition systems in public spaces could enable unprecedented levels of tracking and surveillance, compromising individual privacy and anonymity.

These privacy risks are distinct from other legal effects and can have significant impacts on individuals even when no other direct harm occurs. They also intersect with cybersecurity concerns, as AI systems may introduce new attack vectors for data breaches or become targets themselves due to the valuable data they contain or can generate.

To cater for cybersecurity, data protection and privacy risks, we suggest adding a specific risk dimension along the following lines:

h) Risks associated with cybersecurity, privacy and data protection, including adversarial attacks, data breaches, unauthorized access, or the inference and/or disclosure of personal or sensitive information.

This addition would ensure that cybersecurity, privacy and data protection are given appropriate consideration in assessing the risk level of AI systems, prompting developers and deployers to implement guardrails when the risk level is sufficiently high.

2.2 Waste

The definition of general purpose AI models (GPAI) will capture a significant number of AI systems today, including objectively low risk systems. Since it is proposed that the guardrails should apply to all GPAI models, we believe this will create wasteful compliance activity which we expect is an unintended outcome of the proposal.

GPAI models are defined as “an AI model that is capable of being used, or capable of being adapted for use, for a variety of purposes, both for direct use as well as for integration in other systems.” We contend that this describes most AI systems today, even simple ones.

For example, an AI model built and used internally by a company to predict customer sentiment, satisfaction, or other basic characteristics for the purposes of creating an internal management reporting dashboard could theoretically be used for many other purposes by that company. This appears to satisfy the “capable of being adapted for use” leg of the test above, so would be a GPAI.

Another example is retail sales forecasting models, which aim to predict future units sold based on various input factors. The same models can be (and frequently are) put to use in a range of applications within a retailer, including inventory replenishment, promotion planning, sales strategy, range and assortment planning, and many others. While none of these applications would appear to meet the definition of ‘high-risk’ in isolation, and such models would not typically be thought of by practitioners as ‘general purpose’, they nonetheless fit the proposed definition of GPAI.

There are many similar examples – simple, objectively low risk internal AI systems used within companies can almost always be adapted for secondary uses.

We think this creates unnecessary overlap with the general framework for designating an AI system as ‘high risk’. This could be resolved by considering where and how a model is used. If a model publisher retains complete control over how that model is used (for example if it is used within a

company, with any reuse also being in the company's direct control), then each separate use case can be evaluated under the previous risk framework, and the guardrails applied as needed. There is no need for a separate GPAI designation.

However, if the model is published in a manner where other parties could use, copy or adapt it in ways which might be 'high risk', then the guardrails perhaps ought to apply for that possibility.

With that in mind, we suggest the following alternative definition of GPAI:

"GPAI is an AI model that is intended to be used by or adapted for use by parties separate from the model developer, for a wide range of purposes which may include purposes which would otherwise be designated as 'high-risk'".

A definition of this form would still capture what we believe is the intended target of the proposal – general purpose foundation models or applications based on such models with broad functionality that are exposed to the general public for use or adaptation. However, this definition would not capture narrow but theoretically reusable AI systems operating within companies, under control of the company's staff, where the guardrails and risk assessment process can be considered on a case-by-case basis.

2.3 Confusion

The main aspect likely to create confusion is the absence of a clear severity threshold. The existence of criterion (f) suggests that not all AI systems hitting one or more of criteria (a)-(e) would be considered 'high risk', else (f) is unnecessary.

However, without a well-defined severity threshold, there will be genuine confusion and disagreement about whether an AI system is to be considered 'high risk' or not. While examples and guidance may provide some assistance, the intended breadth of application makes that challenging – novel cases will emerge where guidance is lacking, and the thresholds will then be a matter of significant judgement.

A related aspect likely to cause confusion is the general use of the word 'risk'. Risk requires some uncertainty (per ISO31000: "The effect of uncertainty on..."). Despite the language, proposed principles (a)-(e) tend to describe dimensions of impact that an AI system may have, rather than risks. It should be relatively easy to determine if an AI system could have an effect along those dimensions, but it is likely far more challenging to determine how likely or widespread that impact might be. Principle (f), then, is doing a lot of work, but feels like it would be reliant on a potentially highly subjective forward-looking estimation of the likelihood of potentially low-probability but severe events. The subjectivity emerges even though AI Systems are quantitative in nature, since as a system operates future events will occur which can cause the AI System to deviate from historic performance or predictable patterns.

Many actuaries have experience in conducting such estimates, even subjectively, and we would generally not suggest that such estimates have sufficient reliability to be used to determine whether a threshold of regulation is passed or not, *ex ante*. This uncertainty will lead to inconsistent application – different practitioners will have different opinions about similar AI systems, potentially with entirely good reasons.

3. The Guardrails

We believe that the high-level guardrails represent sensible ‘headlines’ that many people will agree with. However, we do not believe there is yet sufficient detail of the requirements under each proposed guardrail to come to a firm view on what will be required, and whether that is reasonable.

As already noted, we are also generally not convinced that the guardrails should be restricted to AI systems.

As the Institute stated in last year’s consultation, we should “...target risk-management interventions to the risks identified for each situation considered, rather than bluntly applying the same interventions across a broad, vaguely defined risk category.” We said this because such an approach will lead to both gaps and overreach. We identify several examples (below) of gaps and overreach which will occur if the framework is applied as written. The examples should not be taken only as suggestions of things to be fixed, but as an illustration of the structural challenges which will emerge if this approach is pursued. Whatever the content of the guardrails, a blunt, aggregate approach such as that described will lead to waste and gaps. Instead, we suggest a nuanced approach where interventions are imposed in situations where they genuinely add value, and not where they do not.

3.1 Illustrating likely shortcomings with the proposed guardrails

The link between the guardrails and the risks posed by AI systems is unclear. As a result, there are likely to be risks or impacts which are not well managed by the guardrails. One way to illustrate this is to consider the five examples of historical harms outlined on pp.12-13 of the proposals paper and consider whether each of the guardrails would have acted to prevent that harm had they been in place. A brief qualitative analysis of this form is shown in the following table.

Guardrail	AI Resume Screening Discrimination	Facial Recognition Software Bias	Content Moderation Algorithm Bias	AI Misappropriation of First Nations Cultural Material	Educational AI Bias
1. Accountability Process	While biases could arise due to unclear accountabilities, this would not be the primary source of the problem. An accountability process may help to discover issues before they are implemented, but this is not a strong control.	Accountability is generally clear within law enforcement, which tends to operate with strict hierarchies, and yet the examples outlined still occurred. We suggest this control adds little incremental value.	The scale and speed of content moderation makes meaningful accountability difficult, and unlikely to address the problem described.	Accountability processes are likely to be ineffective across international boundaries and may not adequately address complex cultural issues.	Accountability is generally clear within education, which tends to operate with strict hierarchies, and yet the examples outlined still occurred. We suggest this control adds little incremental value.
2. Risk Management Process	Could help identify potential for discrimination but may struggle to address subtle or emerging biases. Effectiveness depends heavily on the skills and abilities of the risk management team, rather than a 'process'.	Could help identify potential for discrimination but may struggle to address subtle or emerging biases. Real possibility that risks are 'accepted' for reasons of operational capacity or efficiency.	The dynamic nature of online content and evolving societal norms make comprehensive risk management challenging. May struggle with context-dependent risks.	Traditional risk management processes may be ill-equipped to handle the nuances of Indigenous Cultural and Intellectual Property ('ICIP'). Global nature of AI development complicates risk management.	Could identify some biases but may struggle with deeply embedded educational inequalities and varying standards across institutions.
3. Data Governance	Strong data governance could significantly improve training data quality, but perfect implementation is unlikely. Historical biases in existing data may persist.	While it could ensure more diverse training data, existing biased datasets may continue to influence outcomes. Practical challenges in creating truly representative datasets will likely remain.	Improved data governance could help, but the subjectivity and cultural specificity of content moderation pose ongoing challenges.	While it could help ensure proper sourcing of cultural material, it may not fully address the complexities of ICIP or prevent unintended misuse. Global nature of AI development again complicates matters, with Australia vying for suitable attention.	Could improve data quality, but systemic educational inequalities may continue to influence outcomes. Standardisation across diverse educational systems is challenging.

Guardrail	AI Resume Screening Discrimination	Facial Recognition Software Bias	Content Moderation Algorithm Bias	AI Misappropriation of First Nations Cultural Material	Educational AI Bias
4. Testing and Monitoring	Rigorous testing could identify many discriminatory outcomes, but some subtle biases may still be missed. The effectiveness of ongoing monitoring often depends on limited resources. Genuine disagreement over what constitutes a bias feels likely.	While testing across diverse populations could reveal biases, real world performance may differ from test conditions. Continuous monitoring in law enforcement contexts may be inconsistent, and incumbent biases may themselves lead to oversights in the collection of future data (due to false negative matches being more prevalent in certain subgroups).	Regular testing could catch many issues, but the volume and variety of online content make comprehensive testing impractical. Evolving cultural norms pose ongoing challenges. Genuine disagreement over what constitutes a bias feels likely.	Testing might identify obvious appropriation, but the vast diversity of cultural expressions makes comprehensive testing nearly impossible.	Could identify many biases, but real world educational outcomes are complex and influenced by many factors outside the AI system. Genuine disagreement over what constitutes a bias feels likely.
5. Human Oversight	Human oversight could catch obvious discrimination but is often subject to the same biases as the original system and may introduce new biases. Scalability and consistency are significant challenges.	While potentially helpful for critical decisions like law enforcement, human oversight is impractical for all uses of facial recognition. Overreliance on AI may lead to reduced scrutiny.	The scale of content moderation makes comprehensive human oversight impractical. Oversight may be inconsistent or subject to the same biases as the AI.	Unless overseen by appropriate cultural experts (which is impractical at scale, especially internationally), human oversight is unlikely to address cultural appropriation issues adequately.	Human oversight could catch obvious biases but is often subject to the same biases as the original system and may introduce new biases. Implementation at scale across educational institutions is challenging.
6. Inform End-users	While it might increase transparency, it does little to prevent discrimination. Job applicants have limited power to challenge or opt out of AI screening processes.	In law enforcement contexts, individuals often are not informed about facial recognition use until after harm has occurred. Does little to prevent misuse or bias.	Users are often unaware of content moderation processes. Informing users does not directly address bias issues and may be ignored in practice.	Informing end-users about the existence of AI does little to address the core issues of cultural appropriation and ICIP in AI development.	While it could increase awareness, students and parents often have limited ability to opt out of or challenge educational AI systems.

Guardrail	AI Resume Screening Discrimination	Facial Recognition Software Bias	Content Moderation Algorithm Bias	AI Misappropriation of First Nations Cultural Material	Educational AI Bias
7. Challenge Processes	While potentially helpful, power imbalances may discourage challenges.	By the time a challenge process is initiated, significant harm (like wrongful arrest) may have already occurred. Effectiveness limited in fast-paced law enforcement scenarios. Existing judicial process would likely meet the requirements of 'a challenge process' for an arrest.	The sheer volume of content moderation decisions makes comprehensive challenge processes impractical. May address individual cases but not systemic issues. Prior examples exist with known limitations (e.g., process for challenging copyright takedowns on large tech platforms).	Challenge processes are likely to be ineffective across international boundaries and may not adequately address or compensate for cultural harm.	Could be effective for individual cases but may not address systemic biases. Power imbalances may discourage challenges. Effectiveness depends on resources available for thorough reviews.
8. Supply Chain Transparency	While it might help identify sources of bias, it does little to directly prevent discrimination. Complex AI supply chains make true transparency challenging.	Transparency in the facial recognition supply chain may identify bias sources but does little to prevent misuse, without further action.	The complexity of content moderation AI and the often-proprietary nature of algorithms limit meaningful transparency. May not address core bias issues.	While it might help track sources of cultural material, it does little to address fundamental ICIP concerns or prevent misuse.	Transparency in educational AI supply chains may identify bias sources but does little to prevent misuse, without further action.
9. Record Keeping	Good for auditing and improvement but does not directly prevent discrimination. Effectiveness depends on quality of analysis and willingness to act on findings.	While helpful for post-incident analysis, it does not prevent initial harm. May be resisted in law enforcement contexts due to liability concerns.	Useful for identifying patterns of bias, but the volume of decisions may challenge comprehensive record-keeping. May not capture context crucial for content decisions.	While it could help track use of cultural material, it does not inherently prevent misuse. Global nature of AI development complicates comprehensive record-keeping.	Could help identify patterns of bias but does not directly prevent it. Effectiveness depends on thorough analysis and willingness to act on findings.
10. Conformity Assessments	Effectiveness of this guardrail is reliant on the effectiveness of the other guardrails, which appear to offer only modest protective value.				

Guardrail	AI Resume Screening Discrimination	Facial Recognition Software Bias	Content Moderation Algorithm Bias	AI Misappropriation of First Nations Cultural Material	Educational AI Bias
Overall Qualitative Assessment	**Modest Protection** While the guardrails appear to offer some protective value, we are unconvinced that they would offer robust protection against this form of harm.	**Limited Protection** The guardrails appear to offer limited protective value against this form of harm, particularly due to the international sourcing of facial recognition systems.	**Limited Protection** The guardrails appear to offer limited protective value against this form of harm, particularly given the scale at which this operates.	**Limited Protection** The guardrails appear to offer limited protective value against this form of harm, particularly for AI Systems developed overseas.	**Modest Protection** While the guardrails appear to offer some protective value, we are unconvinced that they would offer robust protection against this form of harm.

This analysis should not be interpreted too negatively - as we suggested above, we believe the Guardrails represent sensible 'headlines'. While the table indicates that the guardrails appear to offer only limited or modest protective value in these situations, this should not be taken to mean that the guardrails are not worthwhile at all – there will be situations where they are more helpful in reducing risk.

Instead, we suggest this analysis indicates:

- some of the guardrails are more effective in some situations than in others;
- some of the guardrails appear to be of very limited value in some situations, and may impose costs that are not justified by their impact; and
- further guardrails or controls should be considered which might better meet the causes of the risks identified.

Generally, this analysis shows that even across five comparatively easy-to-identify use cases, it is difficult to come up with a unified one-size-fits-all system of guardrails that would be effective for AI Systems in general. This is why the Institute has long advocated for a taxonomy of mitigations that can be selected from to address the specific risks identified for an AI system, rather than applying interventions across the board to a wide range of AI Systems in many different contexts.

It is also not clear how the proposed guardrails could apply in an effective manner to some GPAI systems. Even considering the GPAI systems available in the market today, such as ChatGPT, the following challenges are clear:

- The scale of deployment prevents effective human oversight (guardrail (5)) at the point where a response or piece of advice is delivered (ChatGPT had 100 million weekly active users in 2023[3]), and although one could interpret this to say that the end user is the oversight as they do not have to act on what ChatGPT says, this ignores direct harms to the end users themselves who may be exposed to inappropriate or dangerous content. The scale is also an impediment to effective monitoring;
- The diversity of possible use cases, and difficulty in anticipating potential future use cases, makes comprehensive *ex ante* testing of the system impossible (guardrail (4));
- Data provenance of training data (guardrail (3)) is among the most sensitive of the developers' commercial materials, and the idea of revealing it in any form would likely be met with very strong resistance;
- Detailed record keeping (guardrail (9)) is in direct conflict with the privacy interests of users; and
- The concept of process challenge (guardrail (7)) cannot easily be applied due to the nature of the service.

Hence, although the proposal asks if mandatory guardrails should apply to all GPAI models (current and future), the Guardrails proposed can be shown to be an inadequate mitigant for harms arising from even for the most obvious example, which is the current market-leading GPAI system. As noted in the introduction, while we generally oppose AI-specific regulation we consider that Guardrails for GPAI systems could be appropriate if the definition can be narrowed suitably, but the analysis above suggests that GPAI Guardrails should perhaps differ from any standard set of Guardrails so as to specifically combat the risks identified for GPAI, rather than AI Systems in general.

[3] https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/

3.2 Examples of inappropriate outcomes arising from the guardrails

In some instances, the guardrails are very prescriptive, which could lead to unintended negative side effects when applied at scale. This goes beyond ineffectiveness at controlling risks – in these instances the guardrails create additional harms or costs on society. We illustrate with two examples below, but these are merely illustrations. Again, this indicates the guardrails need a nuanced implementation rather than a blanket implementation.

1. Advanced Driver Assistance Systems (ADAS) in cars, such as collision avoidance and Automated Emergency Breaking (AEB), are common today. Such systems are likely to be classed as 'AI' and would likely be 'high-risk' since they involve health and safety.

   Guardrail (5) says "...real-time human involvement in an AI system may not always be practical and may even make a system less reliable. In such instances, developers should design the system so that a human can review its operations and outputs and reverse a decision if necessary."

   While in some cases the human driver may intervene, these systems are designed to make irreversible driving decisions when, for example, the driver is incapacitated due to a sudden medical emergency. If an ADAS system fails in such a situation and a car accident occurs, it is not possible to reverse that decision, but the guardrail requires this.

   We do not believe the Government intends to cause ADAS systems – which save lives – to be withdrawn from the market, but this would be the logical outcome of the guardrails as proposed.

2. The application of guardrail (6) may lead organisations to add bland notifications to every digital interaction, somewhat analogous to the privacy warnings that have proliferated after the implementation of GDPR.

   This is not only costly for businesses to implement, it adds to the well recognised problem of information overload for consumers, which is harmful to them.

   - When a customer submits an application for insurance and receives an almost instantaneous quote, is it helpful to anyone for that customer to receive a notification that AI was used to make the decision?
   - Should users be informed that emails were placed in their "spam" folder due to AI?
   - Do we need to be reminded that every advertisement and piece of content presented on social media was curated by AI?

   We urge the Government to carefully consider whether this is the intended outcome and consider narrowing the focus of this guardrail to situations where such information would genuinely add value to consumers.

We do not believe these are isolated or unusual examples. AI systems are already widespread and diverse and will get more so. We urge the Government to reflect on every single word in the proposed guardrails, contemplating a very wide range of use cases, and consider whether that guardrail or component of it is something that ought to apply universally.

In many cases, examples of negative side effects will be found. This is a structural problem with the proposal which will be most keenly felt if regulatory option three is chosen.

One way to counteract this problem could be to create a fallback mechanism for high risk AI systems to avoid a particular guardrail requirement, with suitable reasons. There are various potential options for this, such as a list of carved out use cases or areas of the economy, or a mechanism for application for an exemption.

4. The Regulatory Options

The Institute has long put forward a view that broad AI specific regulation could lead to a range of challenges. The content of this proposals paper, and the observations made above in response to it, reinforce that opinion. These challenges make option three incredibly challenging in practice – we believe that gaps, uncertainties and waste are almost certain and could be very costly. These challenges could be avoided, but this would likely introduce complexity and ongoing cost to the proposal in the form of remedial actions like exemptions or carve outs.

We have previously advocated for the review of, and clarification of, existing regulation. Option one is complementary to this and can avoid some of the challenges identified provided some element of judgement is given to primary regulators in their implementation of the guardrails. For example, if a primary regulator determines that a particular guardrail is either excessive or insufficient, that primary regulator should have the ability to react accordingly.

We do not have a strong view on option two, as it will depend on the specifics which are not yet clear. If framework legislation is highly prescriptive, or primary regulation merely 'points' to it to enact it, then all the challenges of option three may still be present. However, if framework legislation merely acts to provide a base level of consistency, but other regulation retains flexibility to avoid poor outcomes, then this is closer to option one in its flexibility and less vulnerable to the risks we have identified.

Whatever option is chosen, we encourage the Government to carefully consider the clarity of the regulation as applied by practitioners like actuaries:

(a) Can practitioners confidently and consistently categorise an AI system as high risk or not?

(b) Can practitioners confidently and consistently describe what they need to do because of an AI system being categorised as high risk?

If the answer to either of these is no, we should challenge whether that is an acceptable outcome and consider what impact that uncertainty or inconsistency could have on the effectiveness of the regime in managing the risks of AI.

We have already identified in this submission several instances where the answer to one or both questions above would be 'no'. This tells us, as practitioners, that the proposal requires more clarity.

The Institute would welcome working further with Government to refine the proposed framework so that practitioners can confidently answer 'yes' to these questions.

We suggest that future versions of the guardrails should be tested with practitioners, using a range of hypothetical examples, to understand if practitioners answer the questions above in a consistent manner. Such evidence would be either useful validation of the regulation's efficacy or would identify areas for improvement. The Institute would strongly support such an evidence-based approach.