Phone: 1300 00 6842
Email: enquiries@ovic.vic.gov.au
PO Box 24274
Melbourne Victoria 3001

Office of the Victorian Information Commissioner

4 October 2024

AI consultation team
Department of Industry, Science and Resources

By email only: aiconsultation@industry.gov.au

Dear consultation team,

Thank you for the opportunity to provide a submission to the Department of Industry, Science and Resources' (**DISR**) consultation on mandatory guardrails for Artificial Intelligence (**AI**) in high-risk settings.

The Office of the Victorian Information Commissioner (**OVIC**) is the primary regulator of information privacy, information security and freedom of information for the Victorian public sector. OVIC's role includes upholding, and advocating for, the privacy rights of the Victorian community, and uplifting information security practice across the Victorian public sector.

It is from these perspectives that OVIC makes the following comments on DISR's proposals paper for mandatory guardrails for AI in high-risk settings.

Prohibiting high-risk AI

1. OVIC holds the view that where the use of an AI system or technology is considered to be high-risk, it should be prohibited.

2. The proposals paper refers to the model of AI regulation under the EU AI Act, which explicitly bans certain AI practices.[1] OVIC agrees with this approach and recommends that the Australian Government follows suit. In addition to those practices identified in the EU AI Act, there may be others that would be appropriate to prohibit in the Australian context.

3. In September 2024, OVIC finalised an investigation relating to the use of ChatGPT by a child protection worker. The investigation found that a significant amount of personal and delicate information[2] had been entered into ChatGPT, resulting in a breach of at least two Information

[1] Article 5, EU AI Act.
[2] The term "delicate information" is used in place of what could, in common usage, be described as "sensitive information". This is because "sensitive information" has a specific definition under the Victorian *Privacy and*

Privacy Principles.[3] OVIC issued a compliance notice to the Department of Families, Fairness and Housing, directing the organisation to prohibit the use of Generative AI text tools in child protection settings. OVIC's investigation report noted that 'child protection, by its nature, requires the very highest standards of care.'[4]

4. It is OVIC's view that the use of AI in instances such as this must be prohibited. Mandatory guardrails, no matter how robust, will likely fail to protect the community from harm.

General comments on mandatory guardrails

5. While OVIC's view is that the use of AI in high-risk settings should be prohibited, mandatory guardrails – such as those outlined in the proposals paper – may provide an appropriate overarching framework for other, lower-risk uses of AI.

6. On their own, guardrails are not sufficient to assist AI developers and deployers to understand what they should do to lower the risk in a particular setting. The guardrails must be accompanied by further guidance, examples of actions that could be taken, and common use cases, so it is clear what developers and deployers should do to reduce the risk of harm to individuals. The guardrails, supporting guidance, examples and use cases should be regularly reviewed and updated, to account for technological advancements and new uses of AI systems.

7. It is unclear from the proposals paper what a developer or deployer should (or should not) do if the risks of a particular AI system or technology cannot be mitigated. The Australian Government must be clear what its expectations are in situations where the risk of harm to individuals remains, even after appropriate guardrails are in place.

8. The proposals paper is also unclear on what the consequences are for a developer or deployer that does not comply with the guardrails. OVIC presumes that further consideration will be given to the kinds of regulatory action and penalties that would be appropriate, as the regulatory model is worked through. For the mandatory guardrails to be effective, the penalties for non-compliance must be clear, consistently applied, and enforceable.

9. Further, OVIC notes that tools used within organisations that utilise AI (such as Microsoft Copilot), are typically role-based. Although mandatory guardrails may exist for high-risk settings, it will be difficult for organisations to manage how their personnel are using these tools. Some tasks for which these tools are used may be considered high-risk (for example, writing a legal document for a court), while others will not (such as scheduling a meeting), but

Data Protection Act 2014, which OVIC administers. What individuals may think of as information that is sensitive to them (for example, information they regard as embarrassing or secret), may not fall within that definition. The term "delicate information" is used to refer to such information.

[3] *Investigation into the use of ChatGPT by a Child Protection worker*, OVIC, September 2024, https://ovic.vic.gov.au/wp-content/uploads/2024/09/DFFH-ChatGPT-investigation-report-20240924.pdf.

[4] *Investigation into the use of ChatGPT by a Child Protection worker*, OVIC, September 2024, page 30, https://ovic.vic.gov.au/wp-content/uploads/2024/09/DFFH-ChatGPT-investigation-report-20240924.pdf.

the same tool will be used for both. Deployers, management, and end-users may face challenges when it comes to distinguishing how tools are used in different contexts. For this reason, AI tools may inevitably be available (and used) in high-risk contexts, and go unnoticed unless challenged. Policy should be clear to identify easy pathways for contestability.

Comments on specific guardrails

10. Guardrail 3: this guardrail requires organisations to ensure they have 'appropriate data governance, privacy and cybersecurity measures in place' to protect AI systems.[5] This guardrail should also refer to broader information security requirements to maintain the confidentiality, integrity and availability of information and systems. Steps need to be taken across all the security domains – information security, ICT security, personnel security, physical security and governance.[6]

11. Guardrail 4: to reduce information privacy and information security risks, non-production data should be used when testing AI models and systems. OVIC recommends a prohibition on the use of production data for testing under guardrail 4.

12. Guardrail 5: this guardrail requires the ability for humans to have oversight of, and intervene at any point in, the AI supply chain and AI lifecycle. In addition to what is outlined in this guardrail, OVIC considers it is important that humans have the ability to turn an AI system off if necessary, and that business continuity allows for the organisation to revert to a human-led approach. Organisations must have an incident response plan in place, that can guide their response in the event that an information privacy or security incident has occurred, or an adversarial attack is mounted on an AI system.[7] The Australian Government may also wish to consider whether it is appropriate for a regulator to have the power to turn off a system or suspend its use.

13. Guardrail 8: this guardrail sets out requirements for transparency with other organisations across the AI supply chain, and notes that 'deployers must report adverse incidents and significant model failures to developers.'[8] OVIC suggests that the guardrail also note the importance of informing other organisations of adverse incidents, such as regulators and independent assessors who have certified the use of the AI model.

---

[5] *Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings*, Department of Industry, Science and Resources, September 2024, page 37.
[6] For further information on the security domains, see the Protective Security Policy Framework and the Victorian Protective Data Security Framework and Standards, at https://www.protectivesecurity.gov.au/about and https://ovic.vic.gov.au/information-security/framework-vpdsf/.
[7] As an example, see a recent adversarial attack which poisoned a Large Language Model and allowed for mass exfiltration of important data. See *Hacker plants false memories in ChatGPT to steal user data in perpetuity*, Ars Technica, 25 September 2024, https://arstechnica.com/security/2024/09/false-memories-planted-in-chatgpt-give-hacker-persistent-exfiltration-channel/.
[8] *Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings*, Department of Industry, Science and Resources, September 2024, page 41.

14. Guardrail 9: OVIC disagrees with the suggestion that small to medium sized businesses should be relieved of the 'compliance burden' associated with recordkeeping. The proposals paper notes that 'organisations must keep and maintain a range of records, including technical documentation, about a high-risk AI system over its lifecycle', but that it may be appropriate for small to medium sized businesses 'to complete less detailed documentation under this guardrail'.[9] Any organisation using an AI system that poses harm to individuals, regardless of its size, should be subject to strict compliance requirements, including recordkeeping. The size of an organisation is not a relevant factor to gauge the risk of non-compliance in an AI setting. Guardrail 10 notes that conformity assessments and certification will rely on records captured under guardrail 9.[10] Without consistency in the way that organisations are expected to create and maintain records, it will not be possible to achieve uniform approaches to guardrail 10.

15. OVIC recommends the Australian Government reconsider this position, and instead mandate the guardrails based on the level of risk posed by the AI system, not the type or size of the organisation. By way of comparison, the Commonwealth *Privacy Act 1988* generally exempts a "small business operator" from having to comply with the Australian Privacy Principles.[11] A small business is defined as one that has an annual turnover of less than $3,000,000. This figure is insignificant when we think about the ease with which AI systems can be accessed, and the potential harms that a small business could do with individuals' personal information. The fixed definition of "small business operator" is problematic, and OVIC cautions against such an approach in this case.

16. Guardrail 10: the proposals paper notes that the conformity assessment required by guardrail 10 is 'an accountability and quality assurance mechanism to verify whether organisations have met their legal obligations prior to deploying a high-risk AI system'.[12] It would be helpful to developers, deployers and end users to understand the types of legal obligations they should consider. Given its remit, OVIC is particularly concerned about information management requirements – for example, secrecy provisions in legislation to which organisations are bound. The Australian Government should consider providing more detailed examples that cover a range of legal obligations that may apply.

17. OVIC suggests that the Australian Government consider creating an additional guardrail relating to decommissioning an AI system. Records retention and disposal, information security, information privacy, and other governance factors will be relevant at the end of the AI lifecycle, and may warrant their own guardrail to ensure that deployers are consistently managing the risks involved once an AI system is no longer being used.

18. OVIC also notes that education is critical to AI being used appropriately, whether in low or high-risk settings. As the proposals paper states, the 'promotion of best practice and

[9] Ibid, pages 41-2.
[10] Ibid, page 42.
[11] Sections 6C and 6D of the *Privacy Act 1988*.
[12] *Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings*, Department of Industry, Science and Resources, September 2024, page 42.

education on how to use AI responsibly are essential'.[13] OVIC encourages the Australian Government to consider how the guardrails could provide for appropriate training and education. Article 4 in the EU AI Act contains such a requirement – that providers and deployers of AI systems take measures to ensure a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf.

Regulatory approach

19. The proposals paper provides three options for mandating the guardrails. OVIC's preference is option 3 – a whole of economy approach. This approach would provide consistency in the way that all developers and deployers, no matter their size or which sector they're from, apply the guardrails. The proposals paper notes that option 3 would create consistent, targeted mechanisms for monitoring and enforcing the guardrails,[14] which is critical. Without regulation and enforceability, there is little incentive for organisations to adhere to the guardrails.

20. OVIC recognises that a new regulatory framework may create some difficulties with identifying and managing existing legislation and frameworks across Australia that already apply to AI (for example, information privacy and security frameworks). It will be important for the Australian Government to work closely with existing regulators and other stakeholders to seek feedback and work through issues.

Thank you again for the opportunity to make a submission. I have no objection to this submission being published by DISR, subject to my signature being removed. I also propose to publish a copy of this submission on OVIC's website.

If you have any questions about the comments in this submission, or would like to discuss these issues further, please contact Adriana Nugent, Assistant Commissioner – Policy, at adriana.nugent@ovic.vic.gov.au.

Yours sincerely

[redacted]

Sean Morrison
Information Commissioner

[13] Ibid, page 5.
[14] Ibid, page 43.