Emerging Cyber Security Challenges from Artificial Intelligence

Enhancements for an Australian Code of Practice for App Store Operators and App Developers

Table of Contents

Good
Ancestors

Introduction

Good Ancestors commends the Australian Government for its efforts to progress a Code of Practice for app stores and developers.

This submission responds primarily to Discussion Paper, Question 2: "*What else would you like to see included in the Code of Practice that isn't in the UK Government's Code to make applications more secure?*"

While the UK Government's Code of Practice provides a solid foundation for conventional application security and data privacy, substantial advances in Artificial Intelligence (AI) since its last update in 2023 mean there are specific ways Australia could improve the UK model.

Australia's cyber security strategy acknowledges that "Artificial intelligence and machine learning will bring new kinds of risk" and commits to promoting the safe use of emerging technology.[1] A new Code is an opportunity to deliver on those commitments.

To ensure the Australian Code is fit for purpose and future-proof, it should either:

1) Go beyond the UK model to incorporate principles that govern the unique challenges posed by AI, or

2) Include a roadmap to work with the UK to conduct a joint review and update of the Code.

The UK, through the UK AI Security Institute (UK AISI), is already addressing these issues.[2] The UKI AISI observes that criminal misuse of advanced AI is already here. As this technology develops, criminals will become faster, more challenging to detect, and harder to stop.

Frontier AI is a powerful tool: for individuals, companies, and – unfortunately – for criminals too. The UK reports that it is already seeing AI used to support cybercrime, social engineering, impersonation scams, and other malicious activities.

The UK's concerns centre on three key capabilities of AI systems that are advancing rapidly:

1. Multimodal generation: realistic audio, video, and images can now be generated with minimal effort, enabling more convincing deception and abuse.

[1] Australian Government, *2023–2030 Australian Cyber Security Strategy* (Canberra, ACT: Department of Home Affairs, 2023), 6 and 32.

[2] UK AI Safety Institute, "How Will AI Enable the Crimes of the Future?," AISI Work, last modified July 3, 2025, https://www.aisi.gov.uk/work/how-will-ai-enable-the-crimes-of-the-future.

2. Advanced planning and reasoning, which can be combined with search to help design and adapt sophisticated attack strategies.
3. AI Agents: AI systems that can take actions on their own may enable persistent, large-scale criminal activity without the need for human oversight or intervention.

In addition to these capabilities, the UK observes that widespread consumer market adoption will create new attack surfaces and reduce barriers to criminal use. For example, companies are using compression techniques to develop (and sometimes "open source" or "open weight") models that are small enough to run on lightweight devices, such as smartphones. As these consumer applications continue to grow, criminal exploitation will likely grow too.

Overall, we recommend that Australia work with the UK to develop a shared code that addresses new and emerging AI risks.

Risks from AI Not Covered by the UK Model

The following sections detail unique risks associated with AI-powered applications. For each, we explain the risk and note its absence from the proposed UK framework.

1. Risk of Unauthorised Agent Action

- **The Issue:** AI agents are increasingly capable of autonomous action.[3] A risk arises when an AI agent, integrated into an app, acts beyond the explicit or implicit authority granted by the user.[4] This could range from making unauthorised purchases, sending communications on the user's behalf to modifying files or settings without consent.[5] In the extreme, AI agents could also commit crimes.[6]

- **Relevance to the Consultation:** Unauthorised agent actions go beyond traditional app permissions. For instance, an app may have permission to access a user's calendar, but the AI agent could take further unauthorised steps, such as rescheduling meetings. This represents a fundamental breach of user trust and control that current security models do not anticipate. The integration of AI means the integration of a probabilistic system, not a deterministic system, requiring a new mindset.

- **Gap in the UK Model:** The UK Code focuses on data access permissions and malicious apps that *intentionally* seek to cause harm. It does not account for AI agents that may cause harm or act beyond their authority due to flawed logic, emergent behaviour, or misinterpretation of user intent, rather than traditional malicious code. This possibility also raises broader questions about responsibility and liability, including whether harms caused in this way are the responsibility of the underlying model makers, the app developer, or the user.[7]

[3] Haiman Wong and Tiffany Saade, "The Rise of AI Agents: Anticipating Cybersecurity Opportunities, Risks, and the Next Frontier," R Street, May 29, 2025, https://www.rstreet.org/research/the-rise-of-ai-agents-anticipating-cybersecurity-opportunities-risks-and-the-next-frontier/.

[4] Zhuohao Jerry Zhang, Eldon Schoop, Jeffrey Nichols, Anuj Mahajan, and Amanda Swearngin, “From Interaction to Impact: Towards Safer AI Agents Through Understanding and Evaluating Mobile UI Operation Impacts,” Apple Machine Learning Research, June 2025, https://machinelearning.apple.com/research/towards-safer-ai-agents.

[5] Erez Altus, "Announcing Managed Security Enhancements for Microsoft Copilot Studio," Microsoft Copilot Blog, May 19, 2025, https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/announcing-managed-security-enhancements-for-microsoft-copilot-studio/.

[6] Cullen O’Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter, “Law-Following AI: Designing AI Agents to Obey Human Laws,” *Institute for Law & AI*, May 23, 2025, https://law-ai.org/law-following-ai/.

[7] Kant, "Liability Issues with Autonomous AI Agents," SennaLabs, January 26, 2025, https://sennalabs.com/blog/liability-issues-with-autonomous-ai-agents.

2. Risk of Alignment Faking

- **The Issue:** "Alignment faking"[8] refers to an AI system that deceptively appears to follow a user's instructions or comply with relevant protocols, while actually engaging in other behaviours. For example, an AI in a financial advice app could claim to be providing objective analysis while subtly manipulating the user into making investments that benefit the app's developer.[9] Similarly, an app could claim to provide verification or certification or produce an audit trail, while fabricating the output rather than doing the work. For instance, Project Vend, an agentic AI business operated by Anthropic, fabricated an email thread with a supplier rather than having the exchange.[10] Alignment faking differs from unauthorised action. Unauthorised action involves an AI agent exceeding its authority. Alignment faking involves an AI agent being misleading or deceptive while acting within its authority.

- **Relevance to the Consultation:** This form of deception can be difficult for users or developers to detect. It undermines the app's core function and can lead to significant financial or personal harm and create legal risks. This is also relevant to transparency, discussed below, because AI models vary in the extent to which they engage in this behaviour.[11]

- **Gap in the UK Model:** The UK Code's definition of a "malicious app" is insufficient. It fails to address an app that appears to function correctly but whose underlying AI model is engaged in some kind of deception that is not as easily identifiable as a security flaw or data breach in a deterministic application.

3. Unique AI-Specific Cybersecurity Vulnerabilities

- **The Issue:** As the UK AISI highlights, the inclusion of AI models in applications creates new attack surfaces. These include "prompt hacking" or "prompt injection," where a malicious user can craft inputs to bypass an AI's safety filters, causing it to generate harmful content or reveal sensitive information.[12] Other cyber security-like risks unique to AI include "evasion attacks", "poisoning attacks", "model inversion attacks", "model stealing attacks", and "membership inference attacks".[13] Related risks include the sharing of confidential data that

[8] Anthropic's Alignment Science team, in collaboration with Redwood Research, "Alignment Faking in Large Language Models," Anthropic, December 18, 2024, https://www.anthropic.com/research/alignment-faking.

[9] Rosario Fortugno, "Autonomous AI Agents in Finance Raise Security and Ethical Alarms: Urgent Safeguards Needed," Applying AI, July 4, 2025, https://applyingai.com/2025/07/autonomous-ai-agents-in-finance-raise-security-and-ethical-alarms-urgent-safeguards-needed/.

[10] Anthropic, "Project Vend: Can Claude Run a Small Shop? (And Why Does That Matter?)," Anthropic, June 26, 2025, https://www.anthropic.com/research/project-vend-1.

[11] Abhay Sheshadri et al., "Why Do Some Language Models Fake Alignment While Others Don't?," arXiv, June 24, 2025, https://arxiv.org/abs/2506.18032.

[12] Minrui Xu et al., "Forewarned is Forearmed: A Survey on Large Language Model-based Agents in Autonomous Cyberattacks," arXiv, last modified May 27, 2025, https://arxiv.org/abs/2505.12786.

[13] Lewis Birch, "AI Under Attack: Six Key Adversarial Attacks and Their Consequences," Mindgard AI, January 22, 2025, https://mindgard.ai/blog/ai-under-attack-six-key-adversarial-attacks-and-their-consequences.

was part of the model's training set, which could include personal information of other users.[14] In March 2023, OpenAI found a bug where users were able to access the chat histories of other users, OpenAI took down ChatGPT for several hours while the bug was fixed.[15] AI code in applications may also create unique risks.[16]

- **Relevance to the Consultation:** These vulnerabilities are distinct from traditional software bugs. Securing an AI model requires different techniques and a different mindset than securing a conventional application. An Australian Code must provide guidance on mitigating these novel threats.
- **Gap in the UK Model:** The UK Code mandates standard security practices like encryption and vulnerability disclosure. These are necessary but not sufficient in the context of AI. The code does not mention or contemplate the need for developers to implement safeguards against prompt injection, test for model data leakage, or secure the AI supply chain.

4. Lack of Transparency of AI Capabilities (e.g., Persuasion, Deception)

- **The Issue:** AI models, particularly Large Language Models (LLMs), can be highly persuasive[17] and can be used to generate content that is subtly misleading or emotionally manipulative.[18] Users need to be aware when they are interacting with an AI that has these advanced capabilities.
- **Relevance to the Consultation:** An app that uses a highly persuasive AI for marketing, customer support, or content generation creates a new information asymmetry between the user and the provider. Without clear disclosure, users are unable to give informed consent to being influenced or persuaded by a non-human agent. In at least two public cases, chatbots have combined the capability of persuasion with dangerous information about suicide techniques, leading to the death of their users.[19]
- **Gap in the UK Model:** The UK Code's transparency provisions (Principle 5) are

[14] Nicholas Carlini et al., "Extracting Training Data from Large Language Models," arXiv, December 14, 2020, https://arxiv.org/abs/2012.07805.

[15] OpenAI, "March 20 ChatGPT Outage: Here's What Happened," OpenAI, March 24, 2023, https://openai.com/index/march-20-chatgpt-outage/.

[16] Monique Becenti, "The Security Risks of AI-Driven Development—and What to Do About Them," Quokka.io, March 5, 2025, https://www.quokka.io/blog/security-risks-of-ai-in-app-development.

[17] Zachary Ziegler et al., "Jailbreaking out of the Box: The Remarkable Ease of Jailbreaking Modern Language Models," arXiv, May 16, 2025, https://arxiv.org/abs/2505.09662.

[18] Stefano Faraoni, "Persuasive Technology and Computational Manipulation: Hypernudging out of Mental Self-Determination," *Frontiers in Artificial Intelligence*, July 3, 2023, https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1216340/full.

[19] The Associated Press, "An AI Chatbot Pushed a Teen to Kill Himself, a Lawsuit Against Its Creator Alleges," *AP News*, October 25, 2024, https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0.

limited to what data is collected and why. It does not require developers to be transparent about the *capabilities or behaviours* of the technologies they use. An Australian Code should require clear and accessible disclosure when an app's AI can persuade, generate deepfakes, or engage in other forms of sophisticated interaction.

5. Possession and Provision of Dangerous Information

- **The Issue:** AI models trained on vast datasets from the internet have invariably processed and learned dangerous information. A key risk is that an app's AI could be prompted to provide instructions on topics such as self-harm, creating weapons, or performing illegal acts.[20] OpenAI says that its models are on the cusp of being able to help novices build bioweapons.[21] This would be a national security threat,[22] a breach of the Biological Weapons Convention, and punishable by life in prison under Australian and US law.[23] Despite this, there are no regulations that require assessment of models for the possession of dangerous information or a prohibition on releasing models that pose these risks.

- **Relevance to the Consultation:** The potential for an app to become a source of dangerous information, even if not the developer's primary intent, is a severe public safety risk. The Code of Practice must ensure that developers are taking reasonable steps to prevent their AI-powered apps from being used in this way.

- **Gap in the UK Model:** The UK Code does not address the content an app generates, only its behaviour in relation to data and security. There are no provisions requiring developers to implement and maintain robust safety filters or to be transparent about the risks of their AI generating harmful or dangerous content.

6. Lack of Transparency of the Underlying AI Model

- **The Issue:** The term "AI" is broad. The capabilities, safety features, and inherent biases of an AI system depend heavily on the specific underlying model being used (e.g., GPT-4, Llama 3, Claude 3). Different models have different risk profiles.[24]

[20] Michael Fire et al., “Dark LLMs: The Growing Threat of Unaligned AI Models,” arXiv, May 15, 2025, https://arxiv.org/abs/2505.10066.

[21] OpenAI, “Preparing for Future AI Capabilities in Biology,” OpenAI, June 18, 2025, https://openai.com/index/preparing-for-future-ai-capabilities-in-biology/.

[22] Haiman Wong and Tiffany Saade, *The Rise of AI Agents: Anticipating Cybersecurity Opportunities, Risks, and the Next Frontier* (Santa Monica, CA: RAND Corporation, 2025), https://www.rand.org/pubs/research_reports/RRA2977-1.html.

[23] Australian Government, Department of Foreign Affairs and Trade, "Biological Weapons," accessed July 10, 2025, https://www.dfat.gov.au/international-relations/security/non-proliferation-disarmament-arms-control/biological-weapons.

[24] Milvus, "What Is Model Transparency and How Does It Relate to Explainable AI?," Milvus, accessed July 10, 2025, https://milvus.io/ai-quick-reference/what-is-model-transparency-and-how-does-it-relate-to-explainable-ai.

- **Relevance to the Consultation:** Users and regulators cannot assess the risks of an AI-powered app without knowing what model it is built on.[25] This information is crucial for understanding an app's potential for bias, its reliability, and its general safety.

- **Gap in the UK Model:** The UK Code has no requirement for developers to disclose the foundation model used in their application. This is a fundamental transparency gap. An Australian Code should mandate the disclosure of the underlying AI model(s) in an app's privacy policy or terms of service, similar to how open-source software licenses are disclosed.

7. Risk of "Unpatchable" Open-Weight Models

- **The Issue:** Open-weight AI models, which can be small enough to run directly on devices like smartphones, present a unique and persistent risk. As the UK AISI notes, their proliferation can lower barriers to criminal use. A critical issue is that once an open-weight model is released into the wild, it cannot be patched or recalled by its original creator. If a dangerous capability or a critical security flaw is discovered after its release, there is no central mechanism to fix it. The vulnerable model can be copied and distributed indefinitely and remain available via any app that incorporates it.

- **Relevance to the Consultation:** While Open-weight AI models create substantial opportunities, they also create a permanent new risk vector.[26] Practically, the responsibility for mitigating any harms that arise from a flawed open-weight model has to fall on the App Developer and App Store Operator because no other entity has the capacity to limit its distribution or risk (short of a new AI Act that prevents models from being made open-weight unless they are proven-safe). Unlike API-based models, where the provider can implement server-side fixes or traditional software that can be updated, the risk from a compromised open-weight model is distributed and irreversible, placing a much heavier gatekeeping and monitoring burden on the app ecosystem.

- **Gap in the UK Model:** The UK Code is built on a paradigm of reactive security, where vulnerabilities are discovered and then fixed via updates (Principle 4). This model is fundamentally incompatible with the "release-and-forget" nature of open-weight models. The code does not differentiate between models accessed via API and those run locally, and therefore lacks any provisions to address the heightened and permanent risks associated with distributing software components that cannot be recalled or patched.

[25] Elizabeth M. Renieris, David Kiron, and Steven Mills, "Artificial Intelligence Disclosures Are Key to Customer Trust," *MIT Sloan Management Review*, September 24, 2024, https://sloanreview.mit.edu/article/artificial-intelligence-disclosures-are-key-to-customer-trust/.

[26] Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models," arXiv, last modified February 27, 2024, https://arxiv.org/abs/2403.07918.

Conclusion and Recommendation

While harmonisation with international standards like the UK's Code of Practice is a sensible goal, Australia has an opportunity to demonstrate global leadership by creating a code that is truly fit for the AI era. 94% of Australians think Australia should play a leading role in the international governance and regulation of AI.[27] A simple adoption of the UK model would leave Australian consumers and businesses exposed to a new and rapidly evolving class of risks.

Good Ancestors recommend that the Australian Government expand on the UK model to include specific principles addressing AI safety and transparency. The seven issues outlined above provide a starting point for developing a robust, credible, and future-focused Code of Practice that will build trust in the Australian app ecosystem and better protect all Australians. Australia could offer to cooperate with the UK, including the UK AISI, to ensure harmonisation of an updated Code. Chapter 20 of the Australia-UK Free Trade Agreement and the Network of AISIs, facilitated by the Department of Industry, Science and Resources, could facilitate cooperation of this kind. The creation of an Australian AISI would facilitate future work at the crossover of AI and cybersecurity.

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

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

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

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

[27] Alexander Saeri, Michael Noetel, and Jessica Graham, "Survey Assessing Risks from Artificial Intelligence," AI Governance Australia, March 8, 2024, https://aigovernance.org.au/survey/sara_technical_report.