Engaging with artificial intelligence

Published by the Australian Signals Directorate's (ASD) Australian Cyber Security Centre (ACSC) on Jan 24, 2024

In partnership with:

- United States (US) Cybersecurity and Infrastructure Security Agency (CISA), the Federal Bureau of Investigation (FBI) and the National Security Agency (NSA)
- United Kingdom (UK) National Cyber Security Centre (NCSC-UK)
- Canadian Centre for Cyber Security (CCCS)
- New Zealand National Cyber Security Centre (NCSC-NZ) and CERT NZ
- Germany Federal Office for Information Security (BSI)
- Israel National Cyber Directorate (INCD)
- Japan National Center of Incident Readiness and Strategy for Cybersecurity (NISC) and the Secretariat of Science, Technology and Innovation Policy, Cabinet Office
- Norway National Cyber Security Centre (NCSC-NO)
- Singapore Cyber Security Agency (CSA)
- Sweden National Cybersecurity Center

Document

https://www.cyber.gov.au/sites/default/files/2025-03/Engaging%20with%20artificial%20intelligence%20%28January%202024%29.pdf

If this link doesn't work, use link in citation

Citation (APA7)

Australian Signalis Directorate: Australian Cyber Security Center. (2024, January 12). Engaging with artificial intelligence. Cyber.gov.au. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/engaging-with-artificial-intelligence

Objective

Educating organizations on the secure use of AI, highlighting important threats and steps to mitigate risks when using AI.

Intended audience

Organizations of all sizes, infrastructure, and governments

Main points

- Focuses on 3 fields of AI: machine learning, natural language processing, and generative AI
- All stakeholders should understand how the threats apply to them and how those threats can be mitigated
- Common AI related threats
    - Data poisoning: the manipulation of training data that results in incorrect or unexpected outputs https://csrc.nist.gov/pubs/ai/100/2/e2023/final
    - Input manipulation or prompt injection: inserting malicious or hidden commands in the input to evade restrictions on the AI
    - Generative AI hallucinations: generative AI producing false or incorrect outputs
    - Privacy and intellectual property concerns: any sensitive information given to AI systems can inform its outputs
    - Model stealing: Using the outputs of an AI to create an approximate replica or to obtain the AI's original training data
- Mitigation considerations
    - Has a **relevant** cybersecurity framework been implemented?
    - How will the AI system use your data (and will it share it)?
    - Does the organization enforce MFA?
    - How will privileged access to the AI system be handled?
    - How will the organization handle backups of the AI system?
    - Can a trial of the AI system be implemented?
    - Is the AI system secure-by-design, including its supply chain?
    - Does the organization understand the limits and constraints of the AI system?
    - Does the organization have suitably qualified staff to ensure the AI system is set-up, maintained and used securely?
    - Does the organization conduct health checks of your AI system?
    - Does the organization enforce logging and monitoring?
    - What happens if something goes wrong with the AI system?