AI
NELSON SANDBOX

Ethical Considerations in Artificial Intelligence

Introduction to Ethical Considerations in AI

- Artificial Intelligence (AI) has become an integral part of our daily lives, revolutionizing industries and shaping societal norms. However, as AI continues to advance, so too do the ethical questions surrounding its development and deployment.

- This document provides an exploration of the ethical landscape of AI, focusing on key areas such as hallucinations, bias, toxicity, and their collective impact on society.

- YouTube link: **Ethics of AI:**

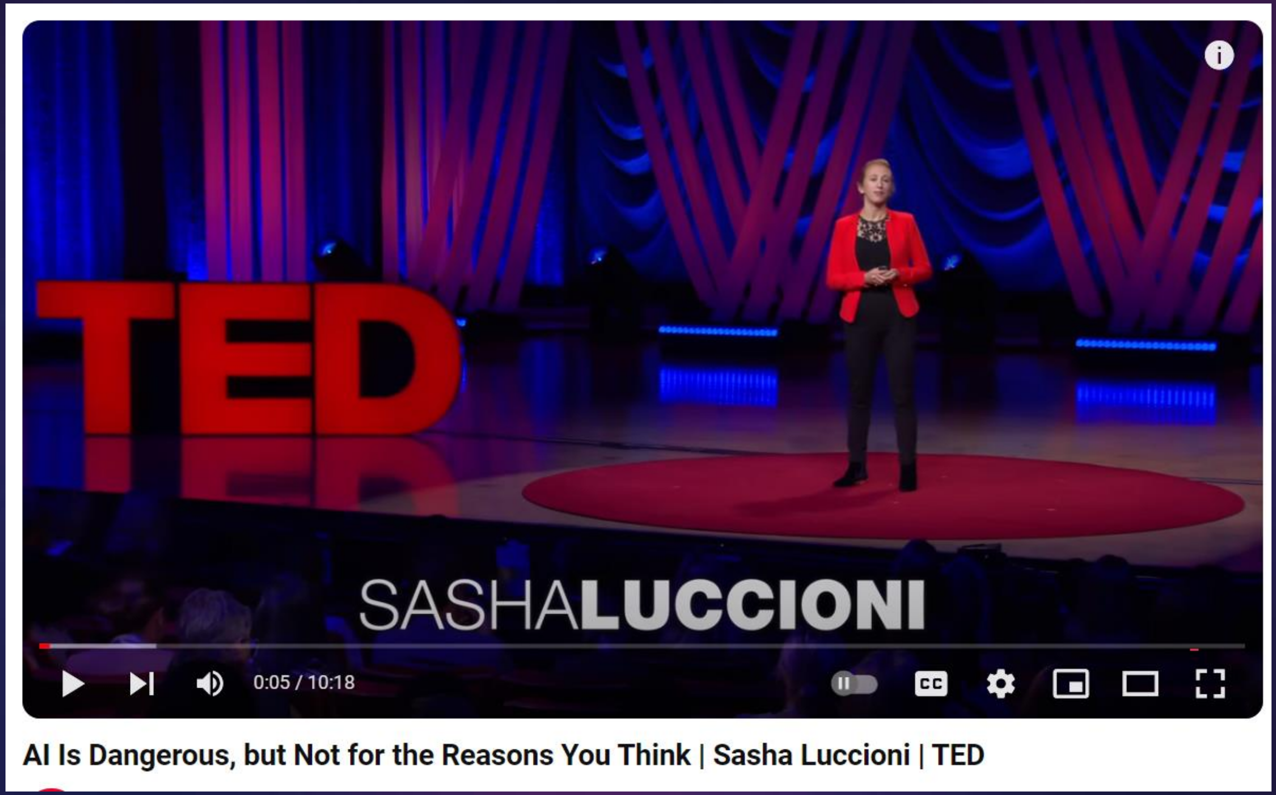
Nelson **AI Sandbox**

General Ethics

- Ethical considerations are the cornerstone of responsible AI development and deployment. At the core of ethical AI lies a commitment to principles such as fairness, accountability, transparency, and responsibility. These principles guide our actions and decisions at every stage of the AI lifecycle, from conception and design to implementation and beyond.
- For example, in the development of AI-powered healthcare systems, ethical considerations ensure patient privacy, data security, and equitable access to healthcare services. Similarly, in the deployment of AI-driven autonomous vehicles, ethical principles govern decisions about safety, liability, and the protection of human life.

By prioritizing ethical conduct in AI development and deployment, we can build trust, foster innovation, and create AI systems that benefit society as a whole.

- YouTube link: **AI Is Dangerous, but Not for the Reasons You Think**



AI Is Dangerous, but Not for the Reasons You Think | Sasha Luccioni | TED

Hallucinations in AI

- Hallucinations in AI, characterized by the generation of false or misleading content, pose significant ethical challenges in the digital age. Deepfake videos, fabricated news articles, and manipulated images are among the deceptive practices that raise concerns about the integrity of information and media authenticity.

- For example, in the realm of politics, deepfake technology has been used to create convincing videos of political leaders making false statements or engaging in inappropriate behaviour, leading to public confusion and misinformation. Similarly, AI-generated text has been employed to fabricate news articles and social media posts, blurring the lines between fact and fiction.

Hallucinations in AI

Examples such as the blurring of lines between fact and fiction underscore the urgent need to address the ethical implications of hallucinations in AI and safeguard the integrity of information in the digital era.

Nelson AI Sandbox

- YouTube link: **Why Large Language Models Hallucinate.**

Why Large Language Models Hallucinate

Large language models (LLMs) like chatGPT can generate authoritative-sounding prose on many topics and domains, they are also prone to just "make stuff up". Literally plausible sounding nonsense! In this video, Martin Keen explains the different types of "LLMs hallucinations“.

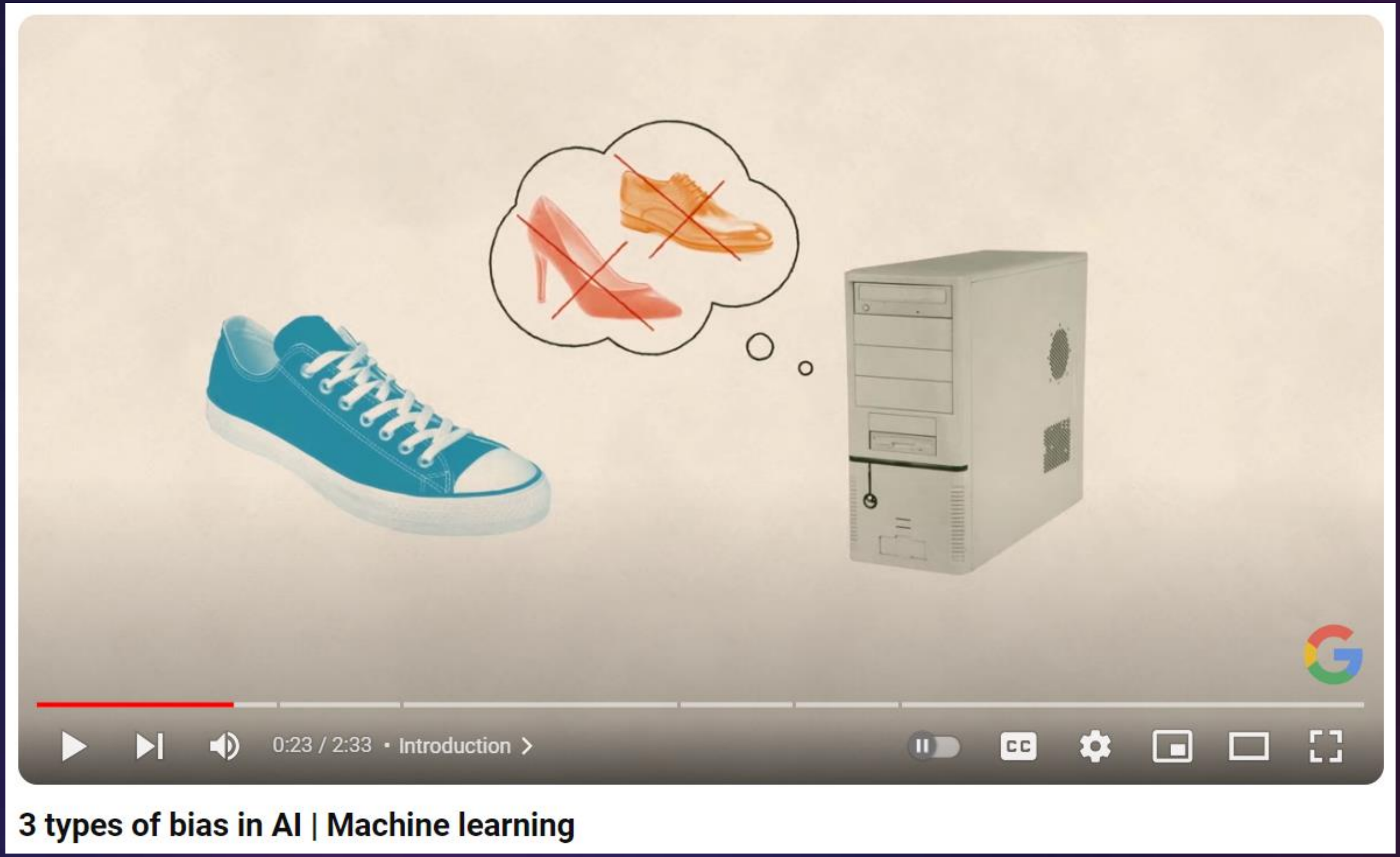
Bias in AI

- Bias in AI systems poses a significant threat to fairness, equity, and social justice. Biases can manifest in AI algorithms and datasets, resulting in discriminatory outcomes across various domains, including criminal justice, hiring, and healthcare.

- For instance, in the criminal justice system, AI algorithms used for risk assessment have been found to exhibit racial bias, leading to disproportionately harsher sentencing for minority defendants. Similarly, in hiring practices, AI-powered recruitment tools have been criticized for perpetuating gender biases and favouring male candidates over female candidates.

Bias in AI

The examples in the previous slide highlight the ethical imperative of addressing bias in AI to ensure fairness and equity for all individuals, regardless of race, gender, or socioeconomic status.



3 types of bias in AI | Machine learning

YouTube link: **3 types of bias in AI.**

Toxicity in AI

- Toxicity in AI refers to the proliferation of harmful content, such as hate speech, misinformation, and online harassment, in digital spaces. The anonymity and reach afforded by online platforms have facilitated the dissemination of toxic content, posing serious risks to individuals' mental health and societal well-being.

- For example, social media platforms like Facebook and Twitter have faced criticism for their failure to effectively moderate toxic behaviour and combat the spread of hate speech and misinformation.

Toxicity in AI

- Additionally, recommendation algorithms on platforms like YouTube have been accused of promoting extremist content, leading to radicalisation and polarisation among users.

- These examples underscore the ethical imperative of addressing toxicity in AI to create safer and more inclusive online environments that foster healthy discourse and respectful interactions.

Collective Impact of Ethics Issues in AI

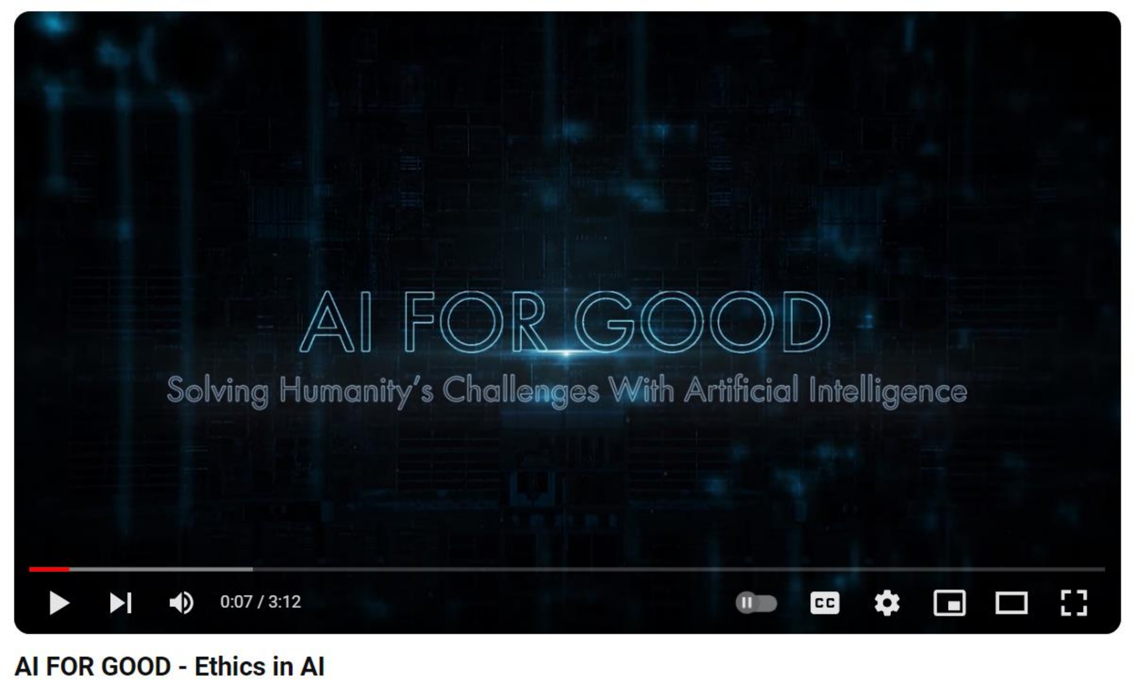
- The collective impact of hallucinations, bias, and toxicity in AI creates a complex web of ethical challenges that demand attention and action. These interconnected issues intersect and compound each other, leading to harmful consequences for individuals, communities, and society as a whole.

- For example, the combination of biased algorithms and toxic content can exacerbate societal divisions and perpetuate harmful stereotypes, further eroding trust in media and information sources, technologists, and stakeholders.

Collective Impact of Ethics Issues in AI

- Additionally, the spread of hallucinated content can undermine democratic values and public discourse, posing significant threats to the integrity of democratic institutions and processes.

- These examples underscore the urgent need for holistic and interdisciplinary approaches to addressing ethics issues in AI, involving collaboration among researchers, policymakers, technologists, and stakeholders.



AI FOR GOOD - Ethics in AI

YouTube link: **AI for good.**

Conclusion

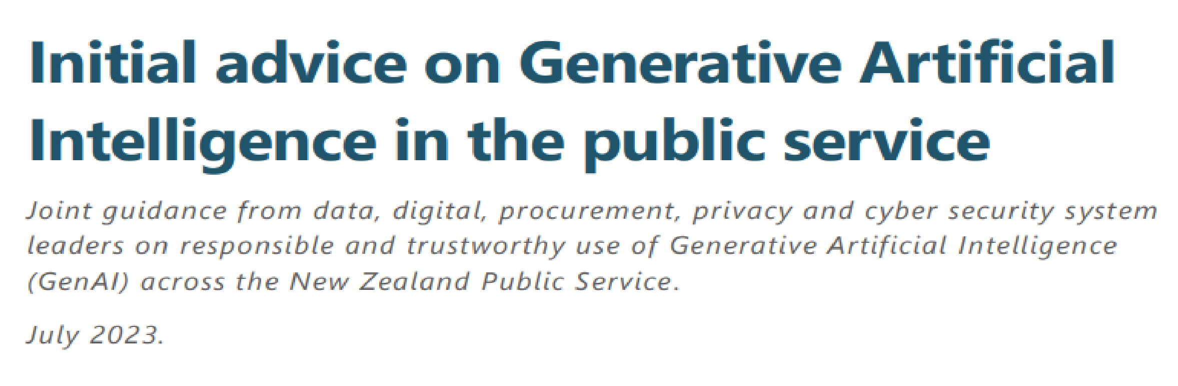
Ethical considerations are central to the responsible development and deployment of AI technologies. By addressing issues such as hallucinations, bias, and toxicity, we can build AI systems that reflect our values and promote the well-being of society.

It is essential for organisations, policymakers, and stakeholders to collaborate in developing ethical frameworks and guidelines to guide the future of AI in a responsible and inclusive manner.

Through collective efforts and ethical leadership, we can harness the transformative potential of AI while mitigating its risks and challenges, paving the way for a brighter and more ethical future.

Useful Reading

NZ Public Service advice:

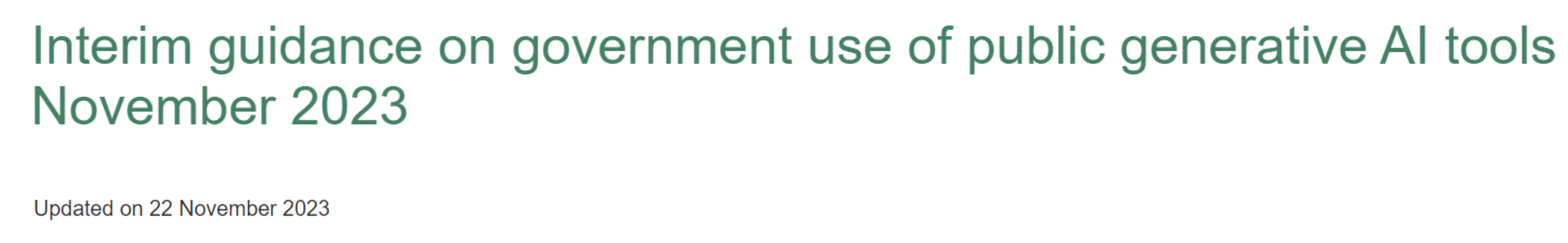


Initial advice on Generative Artificial Intelligence in the public service

Joint guidance from data, digital, procurement, privacy and cyber security system leaders on responsible and trustworthy use of Generative Artificial Intelligence (GenAI) across the New Zealand Public Service.

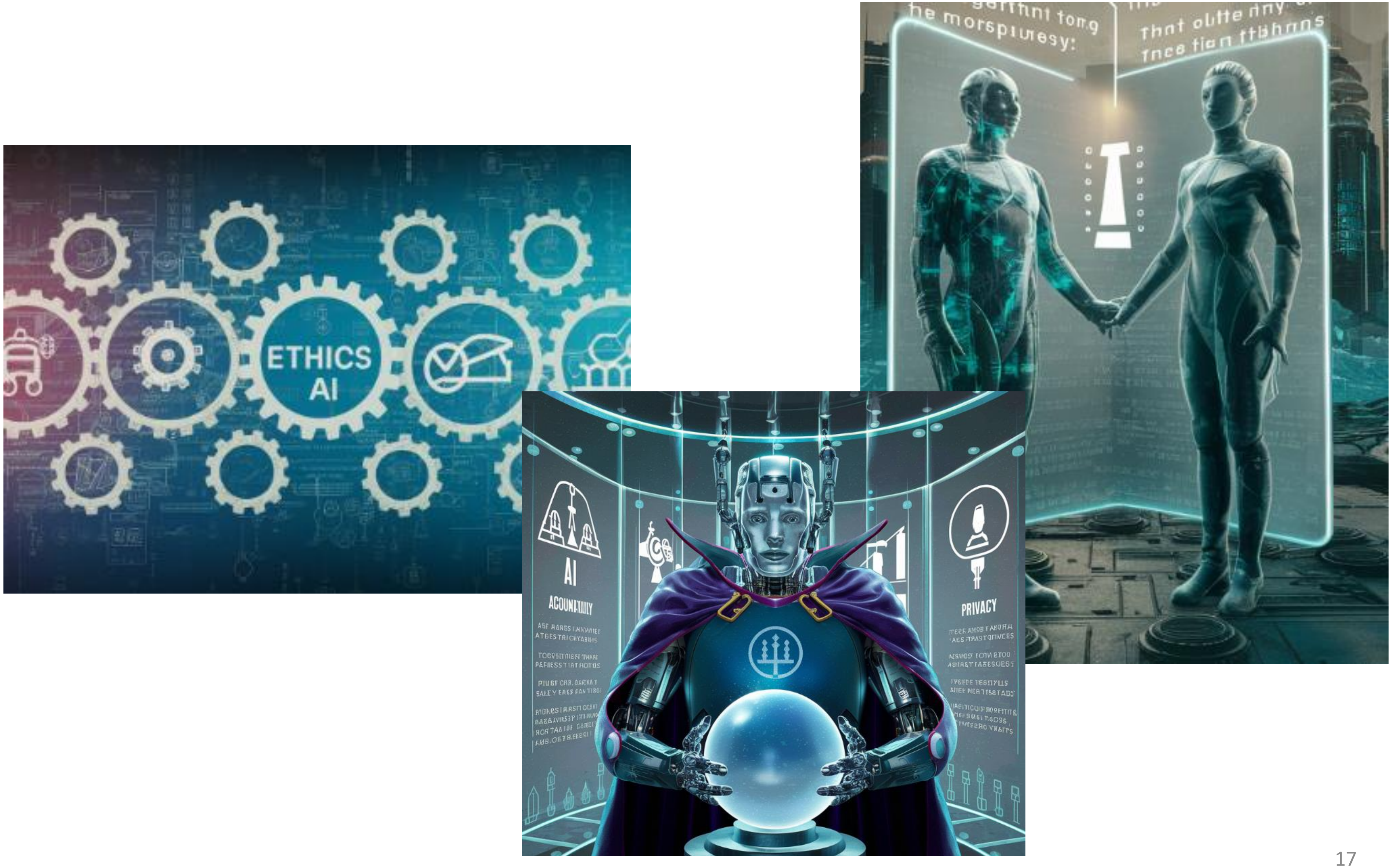
July 2023.

Australian Government Guide:



Interim guidance on government use of public generative AI tools November 2023

Updated on 22 November 2023


ETHICS
AI
AI
ACOUNTAILITY
PRIVACY