Australian Government
Department of Industry, Science,
Energy and Resources

AI Ethics Principles

You can use our 8 principles when designing, developing, integrating or using artificial intelligence (AI) systems to:

- achieve better outcomes
- reduce the risk of negative impact
- practice the highest standards of ethical business and good governance

The principles are voluntary. They are aspirational and intended to complement–not substitute–existing AI related regulations. Read how and when you can apply them (/node/66032).

Principles at a glance

- Human, social and environmental wellbeing: Throughout their lifecycle, AI systems should benefit individuals, society and the environment.
- Human-centred values: Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.
- Fairness: Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.
- Privacy protection and security: Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.
- Reliability and safety: Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.
- Transparency and explainability: There should be transparency and responsible disclosure to ensure people know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them.
- Contestability: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system.

- Accountability: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Principles in detail

Human, social and environmental wellbeing

> Throughout their lifecycle, AI systems should benefit individuals, society and the environment.

This principle aims to clearly indicate from the outset that AI systems should be used for beneficial outcomes for individuals, society and the environment. AI system objectives should be clearly identified and justified. AI systems that help address areas of global concern should be encouraged, like the United Nation's Sustainable Development Goals (https://www.un.org/sustainabledevelopment/sustainable-development-goals/) . Ideally, AI systems should be used to benefit all human beings, including future generations.

AI systems designed for legitimate internal business purposes, like increasing efficiency, can have broader impacts on individual, social and environmental wellbeing. Those impacts, both positive and negative, should be accounted for throughout the AI system's lifecycle, including impacts outside the organisation.

Human-centred values

> Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.

This principle aims to ensure that AI systems are aligned with human values. Machines should serve humans, and not the other way around. AI systems should enable an equitable and democratic society by respecting, protecting and promoting human rights, enabling diversity, respecting human freedom and the autonomy of individuals, and protecting the environment.

Human rights risks need to be carefully considered, as AI systems can equally enable and hamper such fundamental rights. It's permissible to interfere with certain human rights where it's reasonable, necessary and proportionate.

All people interacting with AI systems should be able to keep full and effective control over themselves. AI systems should not undermine the democratic process, and should not undertake actions that threaten individual autonomy, like deception, unfair manipulation, unjustified surveillance, and failing to maintain alignment between a disclosed purpose and true action.

AI systems should be designed to augment, complement and empower human cognitive, social and cultural skills. Organisations designing, developing, deploying or operating AI systems should ideally hire staff from diverse backgrounds, cultures and disciplines to ensure a wide range of perspectives, and to minimise the risk of missing important considerations only noticeable by some stakeholders.

Fairness

> Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

This principle aims to ensure that AI systems are fair and that they enable inclusion throughout their entire lifecycle. AI systems should be user-centric and designed in a way that allows all people interacting with it to access the related products or services. This includes both appropriate consultation with stakeholders, who may be affected by the AI system throughout its lifecycle, and ensuring people receive equitable access and treatment.

This is particularly important given concerns about the potential for AI to perpetuate societal injustices and have a disparate impact on vulnerable and underrepresented groups including, but not limited to, groups relating to age, disability, race, sex, intersex status, gender identity and sexual orientation. Measures should be taken to ensure the AI produced decisions are compliant with anti-discrimination laws.

Privacy protection and security

> Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.

This principle aims to ensure respect for privacy and data protection when using AI systems. This includes ensuring proper data governance, and management, for all data used and generated by the AI system throughout its lifecycle. For example, maintaining privacy through appropriate data anonymisation where used by AI systems. Further, the connection between data, and inferences drawn from that data by AI systems, should be sound and assessed in an ongoing manner.

This principle also aims to ensure appropriate data and AI system security measures are in place. This includes the identification of potential security vulnerabilities, and assurance of resilience to adversarial attacks. Security measures should account for unintended applications of AI systems, and potential abuse risks, with appropriate mitigation measures.

Reliability and safety

> Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.

This principle aims to ensure that AI systems reliably operate in accordance with their intended purpose throughout their lifecycle. This includes ensuring AI systems are reliable, accurate and reproducible as appropriate.

AI systems should not pose unreasonable safety risks, and should adopt safety measures that are proportionate to the magnitude of potential risks. AI systems should be monitored and tested to ensure they continue to meet their intended purpose, and any identified problems should be addressed with ongoing risk management as appropriate. Responsibility should be clearly and appropriately identified, for ensuring that an AI system is robust and safe.

Transparency and explainability

> There should be transparency and responsible disclosure to ensure people know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them.

This principle aims to ensure responsible disclosure when an AI system is significantly impacting on a person’s life. The definition of the threshold for ‘significant impact’ will depend on the context, impact and application of the AI system in question.

Achieving transparency in AI systems through responsible disclosure is important to each stakeholder group for the following reasons[1]

- for users, what the system is doing and why
- for creators, including those undertaking the validation and certification of AI, the systems' processes and input data
- for those deploying and operating the system, to understand processes and input data
- for an accident investigator, if accidents occur
- for regulators in the context of investigations
- for those in the legal process, to inform evidence and decision-making
- for the public, to build confidence in the technology

Responsible disclosures should be provided in a timely manner, and provide reasonable justifications for AI systems outcomes. This includes information that helps people understand outcomes, like key factors used in decision making.

This principle also aims to ensure people have the ability to find out when an AI system is engaging with them (regardless of the level of impact), and are able to obtain a reasonable disclosure regarding the AI system.

Contestability

> When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system.

This principle aims to ensure the provision of efficient, accessible mechanisms that allow people to challenge the use or output of an AI system, when that AI system significantly impacts a person, community, group or environment. The definition of the threshold for 'significant impact' will depend on the context, impact and application of the AI system in question.

Knowing that redress for harm is possible, when things go wrong, is key to ensuring public trust in AI. Particular attention should be paid to vulnerable persons or groups.

There should be sufficient access to the information available to the algorithm, and inferences drawn, to make contestability effective. In the case of decisions significantly affecting rights, there should be an effective system of oversight, which makes appropriate use of human judgment.

Accountability

> Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

This principle aims to acknowledge the relevant organisations' and individuals' responsibility for the outcomes of the AI systems that they design, develop, deploy and operate. The application of legal principles regarding accountability for AI systems is still developing.

Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. This includes both before and after their design, development, deployment and operation. The organisation and individual accountable for the decision should be identifiable as necessary. They must consider the appropriate level of human control or oversight for the particular AI system or use case.

AI systems that have a significant impact on an individual's rights should be accountable to external review, this includes providing timely, accurate, and complete information for the purposes of independent oversight bodies.

Footnote

[1] (#footonote-1) Content based on the Ethically Aligned Design report by IEEE (https://standards.ieee.org/) .