Ethical Frameworks to Navigate Dilemmas in Data-Driven Technologies

In this chapter, we will explore real-world scenarios that require the application of different ethical frameworks. These scenarios include the trolley problem, emergency & disaster scenarios, patient confidentiality in healthcare, & social issues such as AI harassment. The selection of an appropriate framework is crucial in these scenarios, as different frameworks are geared toward different individuals or organizations. It is essential to know which frameworks are suitable & which ones are not for a particular situation.

What Actions to Take in Emergency & Disaster Scenarios?

In the context of using AI in emergency & disaster scenarios, frameworks such as the Universal Guidelines for AI (UGAI) & the Toronto Declaration (TD) can provide ethical guidance to decision-makers. **For instance**, in a pandemic situation where the virus can spread through close contact, contact tracing technology powered by AI can be used to enhance public health efforts. However, this solution can infringe on people's privacy & autonomy, & the UGAI provides guidance on how to ethically design such a system. The UGAI emphasizes using AI to enhance public safety & holds organizations accountable for their actions. It also includes a termination clause that requires the system to be terminated if human control is no longer possible or if it is being used for other purposes.

On the other hand &, the TD centers on the principles of non-discrimination & international human rights & provides a valid framework for balancing the needs of community public health outcomes with individual privacy protection. It advocates for respecting universally accepted human rights principles, which can have universal applicability across cultural & jurisdictional boundaries. However, the lack of robust guidance on privacy concerns can be a drawback of TD.

Both frameworks provide ethical guidance in emergency & disaster scenarios, but the UGAI focuses more on public safety & accountability, while the TD centers on human rights principles. While they have their pros & cons, they can serve as valuable tools for decision-makers looking to use AI ethically & responsibly.

What If a Biased System Is Still Better than Humans?

The use of AI in radiology & medical imaging scans is becoming increasingly widespread due to its ability to provide high levels of accuracy in detecting diseases & other markers that can aid in determining the course of treatment. However, one of the challenges of using such systems is that they are often unable to explain how they arrived at their decision, which raises ethical concerns. The Montreal Declaration for responsible AI (MDR AI) offers guidance in addressing these challenges through its principles of prudence, human autonomy, & responsibility. These principles ensure that potential negative consequences are anticipated, that there is meaningful human control in the use of the system, & that decisions made align with the values of the organization using the system.

While the MDR AI offers a comprehensive approach to addressing ethical dilemmas in the use of AI in medicine, the Beijing AI Principles also provide guidance through their emphasis on informed consent & educational training. The principle of informed consent ensures that patients underst& the risks of being analyzed by an automated system rather than a human doctor, thus increasing the trustworthiness of the process. Educational training, on the other h&, provides doctors with the knowledge to answer questions about the system's limitations & the impacts of its potential failures on patients.

Both sets of principles have their strengths & weaknesses. The MDR AI offers a comprehensive approach that covers all ethical concerns from the perspectives of both doctors & patients. However, its principles may be subject to inconsistent interpretation & require regulatory instruments to back them up. The Beijing AI Principles offer guidance on specific issues such as informed consent & educational training but may lack concreteness in how to implement them.

Overall, the use of AI in radiology & medical imaging scans holds great promise in improving diagnostic accuracy & treatment outcomes. However, it is crucial to ensure that these systems are used ethically & responsibly and prioritize patient autonomy, accountability, & transparency. Both the MDR AI & the Beijing AI Principles provide valuable guidance in addressing ethical dilemmas in the use of AI in medicine.

Should AI Enable Self-Destructive Behavior?

The use of AI can lead to self-destructive behavior, such as compulsive consumption of media on social media platforms. Design patterns like endless scrolling & personalized content can create addictive behavior, sustained by intermittent rewarding through machine learning. This behavior can be detrimental to the user's well-being, relationships, & ability to work. The ethical dilemma is balancing profit motives against providing meaningful control for users. The Universal Guidelines for AI (UGAI) has data quality principles, secret profiling prohibition, & public safety. These principles provide comprehensive coverage of ethical concerns but lack concrete guidelines for practitioners to follow. The Beijing AI principles emphasize informed consent & ethics by design. Informed consent informs users how their data will be used, while ethics by design prevents the use of addictive interfaces & reward systems. This framework empowers designers to take ethical considerations from the earliest stages of development. However, it lacks specificity & may be applied inconsistently across different products & services. The important step is to leverage these principles in the development of AI solutions to prioritize the benefit of people.

What Limits on Persuasive Technology Are Fair?

AI systems integrated into smart toys can be used to persuade children into specific behaviors & beliefs, making them susceptible to manipulation. The OECD AI principles provide guidelines for responsible disclosure, transparency, privacy, & ensuring well-being to help balance the negative consequences that may arise from the use of AI in smart toys. However, while the OECD principles combine responsible disclosure & well-being, they lack demonstrated use cases, making it challenging to evaluate their efficacy. In contrast, the guidelines for untrustworthy AI provide granular guidance for respecting human autonomy, privacy, data governance, & well-being, ensuring persuasive technologies are within acceptable bounds. The pros of the untrustworthy AI guidelines include being more granular, having demonstrated use cases, & being close to legal & regulatory requirements. However, they are written from the perspective of the European legal & regulatory ecosystem, which may not meet all needs expressed in other jurisdictions. By comparing frameworks & principles, we can address issues of persuasiveness in AI.

What If an AI System Is "Emancipated" from Human Oversight?

In today's world, many popular platforms that connect people with each other are subject to regulations & legal requirements that constrain their behavior to a certain extent. However, some of these platforms' algorithms are so complex that they evade human interpretability & are essentially emancipated from human oversight. This emancipation arises from their pervasiveness & controlling a vast amount of activity on the platform.

This raises a dilemma: without meaningful human oversight, it is hard to evaluate if the platforms are promoting well-being & other desirable pro-social outcomes, while at the same time, the scale & complexity of these platforms make this goal hard to achieve in practice. The Beijing AI Principles highlight the control risks & openness & sharing that are particularly relevant in this scenario. The emancipation of algorithmic systems from human oversight presents a control risk, as there is the potential for specification gaming & reward hacking. Openness & sharing can help alleviate these concerns by providing the public with an accountability mechanism to verify that the emancipated system is still behaving within an acceptable balance & not harming human welfare.

The pros of using the Beijing AI Principles are that they explicitly call out control risks, which are the biggest challenge when thinking about the emancipation of automated systems for meaningful human oversight. However, they still lack granular advice, & practitioners might struggle to integrate specific actionable recommendations into their work. On the other h&, the OECD AI Principles advocate for appropriate human intervention when needed, presenting an opposition to the notion of complete emancipation of the system from human oversight & control. The framework provides clear guidelines on how to address the ethical concerns in this scenario, emphasizing the need for maintaining appropriate human controls.

However, the complexity of interactions between the platform & a large user base might be further complicated with the use of opaque black-box AI systems. Under the OECD AI Principles, practitioners are nudged towards using simpler models to limit the emancipation of the AI system from human control. The pros of using the OECD AI Principles are that they are rooted in human rights law & aligned with legal & regulatory requirements, at least in OECD countries. However, their high-level nature creates an additional risk of ethical concerns being raised & open to interpretation, potentially leading to situations where automated systems engage in unethical behavior.

How to Protect Against AI-Driven Harassment?

The power of AI techniques has been demonstrated with the use of systems like deep fakes that can generate convincing fake audio & video segments with someone else's resemblance, making it easier for individuals to harass & defame others. With open-source implementations & easily available online training data, such potential harassment can be mounted against an individual. The Toronto Declaration (TD) is based on international human rights, protecting the physical & mental well-being of individuals, & emphasizing the need for deeper consideration of the second-order effects of open publication, using the lens of international human rights law. The TD framework has the widest possible applicability & can align the largest number of jurisdictions, achieving consistency in deploying such systems worldwide. However, the TD does not provide prescriptive guidance for alternate publishing models & incentives that could encourage open research while protecting people's rights.

The Asilomar AI Principles (AAIP) provide normative guidance on research goals, research funding, research culture, & shared prosperity. The AAIP explicitly addresses concerns on which research problems should be pursued & how to balance downstream impacts on the research community, the target audience, & follow-on research that may spawn from it. The research health of AAIP sets expectations for the problems worth pursuing & those that should not be pursued, with a lens of balancing the costs & benefits of downstream research. The shared prosperity tenant provides grounding for researchers to make trade-offs in understanding when they should work on specific research problems & how to pick the ones that will bring the maximum benefit to the largest number of people.

The AAIP is highly relevant because it addresses ethical concerns related to the pursuit of research & balances the costs & benefits of downstream research. However, they lack use cases that have been implemented in practice, making it hard to assess their efficacy in resolving ethical dilemmas. Borrowing from the field of bioethics may help address some of these concerns, given their experience with similar ethical dilemmas. Overall, the need for frameworks like the TD & AAIP highlights the importance of considering the potential consequences of AI systems, ensuring that they align with international human rights, & promoting open research while protecting people's rights.

Universal Guidelines for Artificial Intelligence

New developments in Artificial Intelligence are transforming the world, from science & industry to government administration & finance. The rise of AI decision-making also implicates fundamental rights of fairness, accountability, & transparency. Modern data analysis produces significant outcomes that have real-life consequences for people in employment, housing, credit, commerce, & criminal sentencing. Many of these techniques are entirely opaque, leaving individuals unaware whether the decisions were accurate, fair, or even about them.

We propose these Universal Guidelines to inform & improve the design & use of AI. The Guidelines are intended to maximize the benefits of AI, to minimize the risk, & to ensure the protection of human rights. These Guidelines should be incorporated into ethical standards, adopted in national law & international agreements, & built into the design of systems. We state clearly that the primary responsibility for AI systems must reside with those institutions that fund, develop, & deploy these systems.

1. **Right to Transparency.** All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, & techniques that produced the outcome.
2. **Right to Human Determination.** All individuals have the right to a final determination made by a person.
3. **Identification Obligation.** The institution responsible for an AI system must be made known to the public.
4. **Fairness Obligation.** Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions.
5. **Assessment & Accountability Obligation**. An AI system should be deployed only after an adequate evaluation of its purpose & objectives, its benefits, as well as its risks. Institutions must be responsible for decisions made by an AI system.
6. **Accuracy, Reliability, & Validity Obligations**. Institutions must ensure the accuracy, reliability, & validity of decisions.
7. **Data Quality Obligation**. Institutions must establish data provenance, & assure quality & relevance for the data input into algorithms.
8. **Public Safety Obligation.** Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices & implement safety controls.
9. **Cybersecurity Obligation**. Institutions must secure AI systems against cybersecurity threats.
10. **Prohibition on Secret Profiling**. No institution shall establish or maintain a secret profiling system.
11. **Prohibition on Unitary Scoring**. No national government shall establish or maintain a general-purpose score on its citizens or residents.
12. **Termination Obligation.** An institution that has established an AI system has an affirmative obligation to terminate the system if human control of the system is no longer possible.

EXPLANATORY MEMORandUM & REFERENCES (https://thepublicvoice.org/ai-universal-guidelines/memo/)

Toronto Declaration

The Toronto Declaration is a set of principles for responsible artificial intelligence (AI) development & deployment, developed by a group of scholars & experts in the field of AI & machine learning (ML). It was presented at the 2018 Conference on Neural Information Processing Systems (NeurIPS) in Toronto, Canada.

The Toronto Declaration aims to highlight the potential risks & negative consequences of AI & promote responsible AI research & development. The Declaration recognizes the power of AI to transform society in positive ways but also acknowledges that the misuse of these technologies can cause harm.

The Declaration outlines four key principles for responsible AI:

1. AI should be designed for the well-being of all humans & the environment: This principle emphasizes the importance of designing AI systems that benefit all humans, including marginalized & vulnerable populations, & that do not harm the environment.

2. AI should operate on principles of transparency, explainability, & intelligibility: This principle stresses the importance of AI systems being transparent, explainable, & intelligible, meaning that they can be understood & audited by humans.

3. Any group or individual involved in the development or deployment of AI should be responsible & accountable for the outcomes of such systems: This principle emphasizes the need for accountability & responsibility throughout the AI development & deployment process.

4. AI should operate within the bounds of human rights, dignity, diversity, & autonomy: This principle highlights the importance of ensuring that AI does not infringe on human rights, dignity, diversity, & autonomy.

The Toronto Declaration also recognizes the importance of interdisciplinary collaboration & the need for ongoing research & dialogue around responsible AI. The Toronto Declaration has been endorsed by hundreds of individuals & organizations in the AI community, including prominent researchers, industry leaders, & civil society groups. It has helped to catalyze discussions & efforts around responsible AI development & deployment & has contributed to the development of guidelines & principles for ethical AI.

Ethics Guidelines for Trustworthy AI

The ethics guidelines for trustworthy AI were developed by the European Commission's High-Level Expert Group on AI & were published in April 2019. These guidelines aim to provide a framework for the development & deployment of trustworthy AI in the European Union.

The guidelines are based on seven key requirements for trustworthy AI:

1. Human agency & oversight: AI systems should empower human beings, allowing them to make informed decisions & to exercise meaningful control over their own lives.

2. Technical robustness & safety: AI systems should be secure, reliable, & resilient enough to operate in a safe & trustworthy manner.

3. Privacy & data governance: AI systems should ensure privacy & data protection, while also promoting the responsible stewardship of data.

4. Transparency: The traceability of AI systems should be ensured in order to allow for accountability & to build trust.

5. Diversity, non-discrimination, & fairness: AI systems should consider the whole range of human abilities, skills, & requirements, & ensure accessibility for all.

6. Societal & environmental well-being: AI systems should be used to enhance positive social change & to improve sustainability & ecological responsibility.

7. Accountability: Mechanisms should be put in place to ensure responsibility & accountability for AI systems & their outcomes.

The guidelines are intended to promote the development of AI systems that are both beneficial & trustworthy, & to foster a European approach to AI that is grounded in the EU's values & principles. The European Commission encourages stakeholders to adopt these guidelines in order to promote the development of trustworthy AI in the EU.

Beijing Artificial Intelligence Principles

The development of Artificial Intelligence (AI) concerns the future of the whole society, all mankind, & the environment. The principles below are proposed as an initiative for the research, development, use, governance & long-term planning of AI, calling for its healthy development to support the construction of a community of common destiny, & the realization of beneficial AI for mankind & nature.

Research & Development:
The research & development (R&D) of AI should observe the following principles:

- **Do Good:** AI should be designed & developed to promote the progress of society & human civilization, to promote the sustainable development of nature & society, to benefit all mankind & the environment, & to enhance the well-being of society & ecology.
- **For Humanity:** The R&D of AI should serve humanity & conform to human values as well as the overall interests of mankind. Human privacy, dignity, freedom, autonomy, & rights should be sufficiently respected. AI should not be used to against, utilize or harm human beings.
- **Be Responsible:** Researchers & developers of AI should have sufficient considerations for the potential ethical, legal, & social impacts & risks brought in by their products & take concrete actions to reduce & avoid them.
- **Control Risks:** Continuous efforts should be made to improve the maturity, robustness, reliability, & controllability of AI systems, so as to ensure the security for the data, the safety & security for the AI system itself, & the safety for the external environment where the AI system deploys.
- **Be Ethical:** AI R&D should take ethical design approaches to make the system trustworthy. This may include, but not limited to: making the system as fair as possible, reducing possible discrimination & biases, improving its transparency, explainability, & predictability, & making the system more traceable, auditable & accountable.
- **Be Diverse & Inclusive:** The development of AI should reflect diversity & inclusiveness, & be designed to benefit as many people as possible, especially those who would otherwise be easily neglected or underrepresented in AI applications.
- **Open & Share:** It is encouraged to establish AI open platforms to avoid data/platform monopolies, to share the benefits of AI development to the greatest extent, & to promote equal development opportunities for different regions & industries.

Use: The use of AI should observe the following principles:

- **Use Wisely & Properly:** Users of AI systems should have the necessary knowledge & ability to make the system operate according to its design, & have sufficient understanding of the potential impacts to avoid possible misuse & abuse, so as to maximize its benefits & minimize the risks.

- **Informed-consent:** Measures should be taken to ensure that stakeholders of AI systems are with sufficient informed consent about the impact of the system on their rights & interests. When unexpected circumstances occur, reasonable data & service revocation mechanisms should be established to ensure that users' own rights & interests are not infringed.
- **Education & Training:** Stakeholders of AI systems should be able to receive education & training to help them adapt to the impact of AI development in psychological, emotional & technical aspects.

Governance: The governance of AI should observe the following principles:

- **Optimizing Employment:** An inclusive attitude should be taken toward the potential impact of AI on human employment. A cautious attitude should be taken toward the promotion of AI applications that may have huge impacts on human employment. Explorations on Human-AI coordination & new forms of work that would give full play to human advantages & characteristics should be encouraged.
- **Harmony & Cooperation:** Cooperation should be actively developed to establish an interdisciplinary, cross-domain, cross-sectoral, cross-organizational, cross-regional, global & comprehensive AI governance ecosystem, so as to avoid malicious AI race, to share AI governance experience, & to jointly cope with the impact of AI with the philosophy of "Optimizing Symbiosis".
- **Adaptation & Moderation:** Adaptive revisions of AI principles, policies, & regulations should be actively considered to adjust them to the development of AI. Governance measures of AI should match its development status, not only to avoid hindering its proper utilization but also to ensure that it is beneficial to society & nature.
- **Subdivision & Implementation:** Various fields & scenarios of AI applications should be actively considered for further formulating more specific & detailed guidelines. The implementation of such principles should also be actively promoted – through the whole life cycle of AI research, development, & application.
- **Long-term Planning:** Continuous research on the potential risks of Augmented Intelligence, Artificial General Intelligence (AGI) & Superintelligence should be encouraged. Strategic designs should be considered to ensure that AI will always be beneficial to society & nature in the future.

Release & Endorsement: The Beijing AI Principles were released on May 25th, Beijing Academy of Artificial Intelligence (BAAI) led the development of the principles, & they have been officially endorsed on the day of release by leading universities (e.g. Tsinghua University, Peking University), & national research institutions (e.g. Institute of Automation, Chinese Academy of Sciences, Institute of Computing Technologies, Chinese Academy of Sciences), & Artificial Intelligence Industry Technology Innovation Strategic Alliance (AITISA), etc.

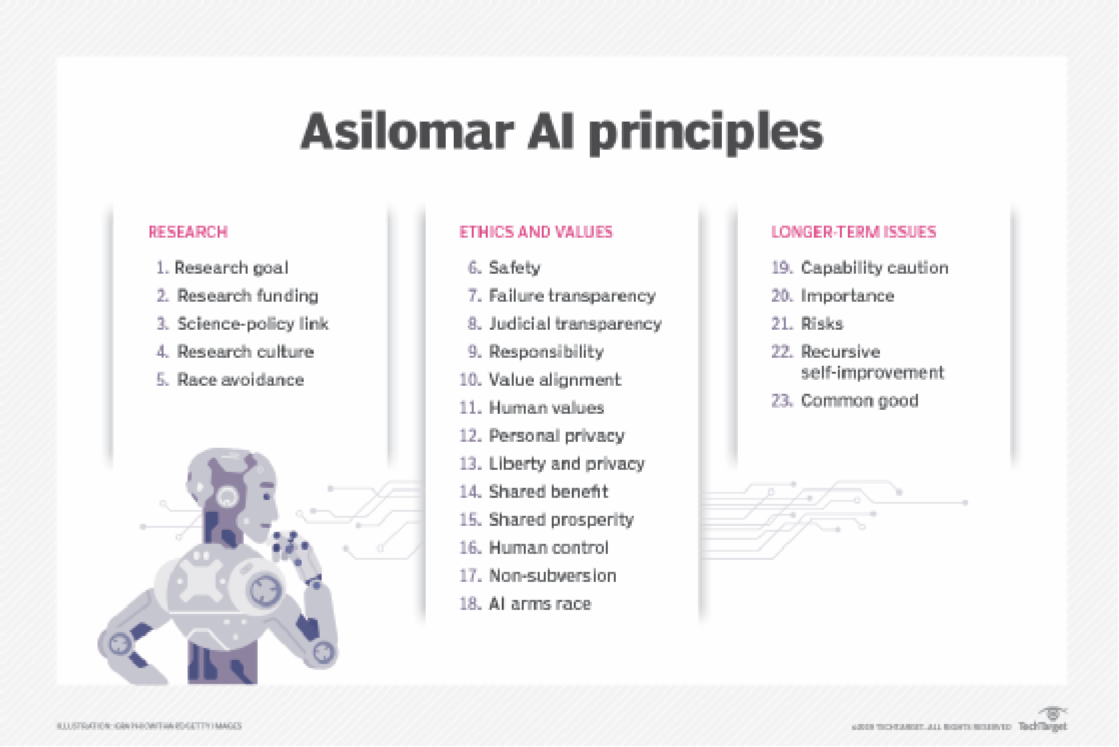
Asilomar AI Principles

Asilomar AI Principles are 23 guidelines for the research and development of artificial intelligence (AI). The Asilomar Principles outline developmental issues, ethics and guidelines for the development of AI, with the goal of guiding the development of beneficial AI. The tenets were created at the Asilomar Conference on Beneficial AI in 2017 in Pacific Grove, Calif. The conference was organized by the Future of Life Institute.

The 23 principles were developed by a group of AI researchers, robotics, technology experts and legal scholars from different universities and organizations. These experts organized the AI principles at the Asilomar Conference while discussing the future of AI and its regulation.

The Future of Life Institute is a nonprofit organization founded in 2014 by MIT cosmologist Max Tegmark, Skype co-founder Jaan Tallinn, physicist Anthony Aguirre, DeepMind research scientist Viktoriya Krakovna and Tufts University postdoctoral scholar Meia Chita-Tegmark. Thousands of AI and robotics researchers have signed onto the principles, as well as with other endorsers from a variety of AI research leaders, including Google, Apple and OpenAI. In 2018, the state of California endorsed these principles.

The Asilomar AI Principles are divided into three categories: Research, Ethics and values, and Longer-term issues. Often, the principles are a clear statement of possible undesirable outcomes, followed by recommendations to prevent such an event.



Asilomar AI principles

RESEARCH

1. Research goal
2. Research funding
3. Science-policy link
4. Research culture
5. Race avoidance

ETHICS AND VALUES

6. Safety
7. Failure transparency
8. Judicial transparency
9. Responsibility
10. Value alignment
11. Human values
12. Personal privacy
13. Liberty and privacy
14. Shared benefit
15. Shared prosperity
16. Human control
17. Non-subversion
18. AI arms race

LONGER-TERM ISSUES

19. Capability caution
20. Importance
21. Risks
22. Recursive self-improvement
23. Common good

ILLUSTRATION: IGBN PHOWTHAROGETTY IMAGES

©2019 TECHTARGET. ALL RIGHTS RESERVED TechTarget

The 23 Asilomar AI Principles are separated into three different categories: Research, Ethics and values, and Longer-term issues.

Research

This subsection of five principles revolves around how AI is researched and developed as well as its transparency and its beneficial use:

1. **Research.** The goal of AI research should be to create not undirected intelligence but beneficial intelligence. This means AI research should always be beneficial.
2. **Research funding.** Investments in AI should be accompanied by funding for research on ensuring its beneficial use.
3. **Science-policy link.** There should be constructive and healthy exchanges between AI researchers and policymakers.
4. **Research culture.** A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
5. **Race avoidance.** Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and values

This subsection of 13 AI principles revolves around the ethics of AI and the values instilled while developing it:

1. **Safety.** AI systems should be safe and secure throughout their operational lifetime and verifiably so where applicable and feasible.
2. **Failure transparency.** If an AI system causes harm, it should be possible to ascertain why.
3. **Judicial transparency.** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
4. **Responsibility.** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse and actions, with a responsibility and opportunity to shape those implications.
5. **Value alignment.** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
6. **Human values.** AI systems should be designed and operated to be compatible with ideals of human dignity, rights, freedoms and cultural diversity.
7. **Personal privacy.** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
8. **Liberty and privacy.** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
9. **Shared benefit.** AI technologies should benefit and empower as many people as possible. This, for example, includes the use of AI to make jobs easier, the optimization of energy use or as expert knowledge-based systems.
10. **Shared prosperity.** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
11. **Human control.** Humans should choose how and whether to delegate decisions to AI systems to accomplish human-chosen objectives.

12. **Non-subversion.** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
13. **AI arms race.** An arms race in lethal autonomous weapons should be avoided.

Longer-term issues

This subsection of five AI principles revolves around the importance, risks and potential good AI can provide in the long term:

1. **Capability caution.** There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
2. **Importance.** Advanced AI could represent a profound change in the history of life on Earth and should be planned for and managed with commensurate care and resources.
3. **Risks.** Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
4. **Recursive self-improvement.** AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
5. **Common good.** Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

OECD AI Principles

The OECD (Organisation for Economic Co-operation and Development) principles for AI are a set of guidelines developed by the OECD to help promote the responsible development and use of AI. The principles were adopted by the OECD Council on 22 May 2019 and are composed of five values-based principles and five recommendations for public policy and international cooperation.

The five values-based principles are:

1. Inclusive growth, sustainable development, and well-being: AI should contribute to inclusive growth, sustainable development and well-being of all people.
2. Human-centered values and fairness: AI should respect human rights, diversity, and the principles of fairness and non-discrimination.
3. Transparency and explainability: AI should be transparent and explainable in order to foster trust and accountability.
4. Robustness, security and safety: AI systems should be robust, secure and safe throughout their entire lifecycle.
5. Accountability: Actors involved in the development, deployment and operation of AI systems should be held accountable for their proper functioning.

The five recommendations for public policy and international cooperation are:

1. Foster investment in research & development for trustworthy AI.
2. Foster an enabling policy environment for AI.
3. Shape an appropriate legal and policy framework.
4. Facilitate public understanding and awareness of AI.
5. Foster international cooperation on AI.