MAPPING PRINCIPLES OF ARTIFICIAL INTELLIGENCE

AUTHORS:
Caroline Burle and Diogo Cortiz

SUMMARY

ceweb.br nic.br cgi.br

Introduction

This is a non-exhaustive multisectoral mapping of Artificial Intelligence principles. We mapped six international initiatives, two from the government sector (European Commission and US Department of Defense), two from the private sector (Google and Microsoft), one international organization (Organization for Economic Co-operation and Development - OECD) and one composed of academy and private sector (Beijing Academy of Artificial Intelligence).

We selected these initiatives in order to ensure regional plurality (United States, Europe and Asia) - no document produced for Brazil was found at the time of writing - and multisectoral (private sector, academia, governments and international organization).

We analyzed the principles of Artificial Intelligence found in each of these initiatives, based on six dimensions: Fairness; Reliability & Safety; Social Impact; Accountability; Privacy & Security; and Transparency.

The six dimensions were elaborated by the authors of this preliminar document, based on the previous readings of each document. After finding similarities between the documents, we realized that these dimensions would be adequate to map the Artificial Intelligence principles of the analyzed initiatives.

Principles of Artificial Intelligence by Dimensions

Fairness

European Commission

The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity in terms of access to education, goods, services and technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives. Seek effective redress against decisions made by AI systems and by the humans operating them.

The entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

The United States Defense Innovation Board

The Department of Defense should take deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons.

Beijing AI Principles

AI Research & Development 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 and biases, improving its transparency, explainability, and predictability, and making the system more traceable, auditable and accountable.

OECD (Organisation for Economic Co-operation and Development)

AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.

Google

Avoid creating or reinforcing unfair bias. AI algorithms and datasets can reflect, reinforce, or reduce unfair biases. We recognize that distinguishing fair from unfair biases is not always simple, and differs across cultures and societies. We will seek to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief.

Microsoft

AI systems should treat all people fairly.

Reliability & Safety

European Commission

The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both

benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity in terms of access to education, goods, services and technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives.
seek effective redress against decisions made by AI systems and by the humans operating them.
the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

The United States Defense Innovation Board

The Department of Defense AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use. The Department of Defense AI systems should be designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behavior.

Beijing AI Principles

Continuous efforts should be made to improve the maturity, robustness, reliability, and controllability of AI systems, so as to ensure the security for the data, the safety and security for the AI system itself, and the safety for the external environment where the AI system deploys.

OECD (Organisation for Economic Co-operation and Development)

AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society. AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.

Google

Be built and tested for safety. We will continue to develop and apply strong safety and security practices to avoid unintended results that create risks of harm. We will design our AI systems to be appropriately cautious, and seek to develop them in

accordance with best practices in AI safety research. In appropriate cases, we will test AI technologies in constrained environments and monitor their operation after deployment.

Microsoft

AI systems should perform reliably and safely.

Social Impact

European Commission

The fundamental rights upon which the EU is founded are directed towards ensuring respect for the freedom and autonomy of human beings. Humans interacting with AI systems must be able to keep full and effective self- determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work.

The United States Defense Innovation Board

Not included.

Beijing AI Principles

AI should be designed and developed to promote the progress of society and human civilization, to promote the sustainable development of nature and society, to benefit all humankind and the environment, and to enhance the well-being of society and ecology. The development of AI should reflect diversity and inclusiveness, and be designed to benefit as many people as possible, especially those who would otherwise be easily neglected or underrepresented in AI applications. It is encouraged to establish AI open platforms to avoid data/platform monopolies, to share the benefits of AI development to the greatest extent, and to promote equal development opportunities for different regions and industries.

OECD (Organisation for Economic Co-operation and Development)

AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.

Google

Be socially beneficial. Uphold high standards of scientific excellence. AI also enhances our ability to understand the meaning of content at scale. We will strive to make high-quality and accurate information readily available using AI, while continuing to respect cultural, social, and legal norms in the countries where we operate. Will continue to thoughtfully evaluate when to make technologies available on a non-commercial basis.

Microsoft

AI systems should empower everyone and engage people.

Accountability

European Commission

Accountability: including auditability, minimisation and reporting of negative impact and trade-offs. The requirement of accountability complements the above requirements, and is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.

The United States Defense Innovation Board

Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of The Department of Defense AI systems.

Beijing AI Principles

Researchers and developers of AI should have sufficient considerations for the potential ethical, legal, and social impacts and risks brought in by their products and take concrete actions to reduce and avoid them.

OECD (Organisation for Economic Co-operation and Development)

Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

Google
Be accountable to people. Design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. AI technologies will be subject to appropriate human direction and control.

Microsoft
AI systems should have algorithmic accountability.

Privacy & Security

European Commission
Includes respect for privacy quality and integrity of data, and access to data. Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

The United States Defense Innovation Board
Not included.

Beijing AI Principles
The Research & Development of AI should serve humanity and conform to human values as well as the overall interests of humankind. Human privacy, dignity, freedom, autonomy, and rights should be sufficiently respected. AI should not be used to against, utilize or harm human beings.

OECD (Organisation for Economic Co-operation and Development)
Not included.

Google
Incorporate privacy design principles in the development and use of AI technologies. Will give opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data.

Microsoft
AI systems should be secure and respect privacy.

Transparency

European Commission

Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights.

The United States Defense Innovation Board

The Department of Defense AI engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation.

Beijing AI Principles

AI Research & Development 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 and biases, improving its transparency, explainability, and predictability, and making the system more traceable, auditable and accountable.

OECD (Organisation for Economic Co-operation and Development)

There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.

Google

Fornecer transparência apropriada e controle do uso de dados. Este princípio é citado dentro de Privacidade & Segurança.

Microsoft

AI systems should be understandable.

Some considerations

The European Commission and the Beijing AI Principles address the six principles listed in this mapping, and the two initiatives have detailed each Artificial Intelligence principle analyzed. Microsoft also addresses all of the principles listed, but it is quite concise without detailing them.

In this regard, it is interesting to note that the European Commission, the politically independent executive body of the European Union, is made up of a team of 28 Commissioners (one from each EU country). Beijing AI Principles is made up of Chinese universities and companies. Therefore, it is analyzed that the two initiatives are composed of several entities, which may have brought greater richness and detail to the principles.

The principle of fairness is addressed by the six initiatives analyzed. The European Commission brings more detail about this principle and proposes a division between substantive and procedural equity. Affirms at first sight the commitment to ensure a fair and equitable distribution of benefits and costs, as well as ensuring that individuals and groups are free from unfair prejudice, discrimination and stigmatization, while the second seeks effective redress against decisions made by intelligence systems. Artificial and the humans who operate them.

Still on fairness, the Beijing Academy of Artificial Intelligence suggests taking ethical design approaches to make the system reliable. The OECD adds that AI systems should include appropriate safeguards - for example, enabling human intervention where necessary - to ensure a just society.

Regarding the Reliability & Safety principle, the six initiatives analyzed explain this principle. All claim that Artificial Intelligence systems must have reliable and reliable performance. The European Commission details that they must protect human dignity as well as mental and physical integrity. In addition to emphasizing that AI systems and the environments in which they operate must be safe, technically robust and must be ensured that they are not open to malicious use. The US Department of Defense adds that the security, protection, and robustness of such systems must be tested and guaranteed throughout the entire life cycle of this domain of use.

The Social Impact dimension is analyzed by five organizations, only the US Department of Defense does not specify this principle. The European Commission states that AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or group humans. Instead, they should be designed to increase, complement and enable human cognitive, social and cultural skills.

The Beijing AI Principles adds on the Social Impact that AI must be designed and developed to promote the progress of society and human civilization, to promote the sustainable development of nature and society, to benefit all humanity and the environment. improve the well-being of society and ecology. Google says it is necessary to provide accurate and high quality information using AI in order to comply with cultural, social and legal standards in the countries in which it operates.

All mapped initiatives cite the Accountability principle. In general, they state that mechanisms need to be put in place to ensure the accountability and accountability of AI systems and their outcomes, before and after their development, deployment and use. The Beijing AI Principles adds that AI researchers and developers should take sufficient account of the possible ethical, legal and social impacts and risks brought about by their products and take concrete actions to reduce and prevent them.

Of the six initiatives mapped, only four spell out the Privacy & Security principle. The European Commission states that it includes respect for the quality of privacy and data integrity and access to data. In addition to being closely linked to the principle of damage prevention. The Beijing Academy of Artificial Intelligence adds that privacy, dignity, freedom, autonomy and human rights must be sufficiently respected. Google says it must incorporate privacy principles in the development and use of our AI technologies. It is Microsoft that Artificial Intelligence systems must be secure and respect privacy.

About the Transparency principle, only Google does not specify, but mentions within Privacy & Security. Overall, initiatives state that processes need to be transparent. The European Commission emphasizes that the capabilities and purpose of AI systems should be openly communicated and decisions - as far as possible - explainable to those affected directly and indirectly. The US Department of Defense adds that they should include transparent and auditable methodologies, data sources and procedures, and project documentation. It is the Beijing AI Principles that ethical design approaches need to be taken to make the system reliable.

We therefore find that three principles: Equity, Reliability & Safety and Accountability are addressed by the six mapped initiatives. The principles of Social Impact and Transparency are spelled out by five initiatives. And only the Privacy & Security principles are detailed by four of the six mapped initiatives.

References

In alphabetical order by initiative

AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense
Defense Innovation Board
https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

Beijing AI Principles
Universidades Chinesas e setor privado
https://www.baai.ac.cn/blog/beijing-ai-principles

Building Guidelines for Trustworthy AI
European Commission (High Level Group on Artificial Intelligence)
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

Google AI Principles
Google
https://www.blog.google/technology/ai/ai-principles/

Microsoft AI Principles
Microsoft
https://www.microsoft.com/en-us/ai/our-approach-to-ai

OECD (Organisation for Economic Co-operation and Development) Principles on AI
OECD (Organisation for Economic Co-operation and Development)
https://www.oecd.org/going-digital/ai/principles/