OECD publishing

SCOPING THE OECD AI PRINCIPLES

DELIBERATIONS OF THE EXPERT GROUP ON ARTIFICIAL INTELLIGENCE AT THE OECD (AIGO)

OECD DIGITAL ECONOMY PAPERS
November 2019 **No. 291**

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This document presents the work conducted by the Expert Group on Artificial Intelligence at the OECD (AIGO) to scope principles to foster trust in and adoption of artificial intelligence (AI), as requested by the Committee on Digital Economy Policy (CDEP). The work was developed over four in-person meetings and several teleconference calls in-between those meetings. The group concluded its discussion and agreed on this draft at its fourth and last meeting in Dubai, UAE, on 8-9 February.

This paper was approved and declassified by the CDEP on 1 July 2019 and prepared for publication by the OECD Secretariat. The description of what is an AI system and the AI system lifecycle informed the CDEP's discussion of a draft Recommendation of the Council on Artificial Intelligence on 14-15 March 2019. The OECD Council adopted this Recommendation at its 1397th Session on 22 May 2019.

This document was a contribution to IOR 1.3.1.1.3 Artificial Intelligence of the 2019-2020 Programme of Work of the CDEP. For more information, please visit www.oecd.ai.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CDEP(2019)1/FINAL

Table of contents

Figures

Background

In the context of its work on Artificial Intelligence (AI), the Committee on Digital Economy Policy (CDEP) agreed, at its meeting on 16-18 May 2018, to form an expert group on AI to scope principles to foster trust in and adoption of AI in society, in view of developing a Council Recommendation in the course of 2019 [DSTI/CDEP/M(2018)1, Item 10].

The group, AIGO, comprised over 50 experts from different disciplines and different sectors (government, industry, civil society, academia and the technical community; see also Annex B: List of AIGO members). AIGO held four meetings: two meetings in Paris in September and November 2018, one at MIT in January 2019, and a last meeting in Dubai, early February 2019.

Chaired by the CDEP Chair, Mr Wonki Min[1], the group's objective was to scope principles with the following characteristics: specific to AI, facilitating innovation and trust in AI, implementable, flexible to stand the test of time, and conducive to increased co-operation. At each meeting, the group discussed proposals for the principles, revised by the Secretariat based on oral input from the previous discussion and on written input. The group also formed two subgroups, to discuss and clarify particular technical aspects, namely, articulating a common understanding of "AI systems" (Chapter 1) and of the AI system lifecycle (Chapter 2). The work benefited from the diligence, engagement and substantive contributions of its members, as well as from their multi-stakeholder and multidisciplinary backgrounds.

At its meeting in Dubai on 8-9 February 2019, the group agreed on its final proposal to the Committee, which included five value-based principles that AI should promote, four recommendations for national AI policies, and a principle on international cooperation for trustworthy AI (Annex A). These principles aim to apply globally to all stakeholders and throughout the entire AI life cycle.

The group's proposal was submitted to the Committee to inform its discussion of a draft Recommendation of the Council on Artificial Intelligence on 14-15 March 2019. It was subsequently adopted by the OECD Council at Ministerial level on 22 May 2019.

What is an 'AI system'

In November 2018, AIGO set up a subgroup to develop a description of an AI system, in view of delineating the scope of applicability of the OECD Principles. This chapter details the high-level description of an AI system provided in the Principles. The description aims to be understandable, technically accurate, technology-neutral, and applicable to short and long-term time horizons. It is broad enough to encompass many of the definitions of AI commonly used by the scientific, business and policy communities.

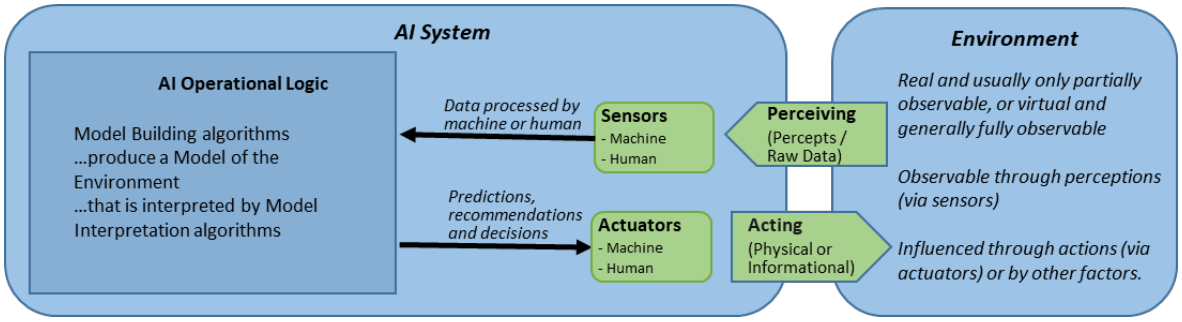
Twenty-one AI experts participated in the work of the subgroup, which was co-moderated by Mr. Marko Grobelnik from Slovenia and by Mr. Javier Juarez Mojica from Mexico. Mr. Marko Grobelnik authored the present document with input from the subgroup that met regularly from mid-December 2018 to mid-February 2019 and from the Secretariat.

Conceptual view of an AI system

The present description of what is an AI system is based on the conceptual view of AI detailed in "Artificial Intelligence: A Modern Approach" (Russel, S. & Norvig, P., 2009[1]). This view is consistent with a widely-used definition of AI as "the study of the computations that make it possible to perceive, reason, and act" (Winston, 1992[2]) and with similar general definitions (Gringsjord, S. & Govindarajulu, N.S., 2018[3]).

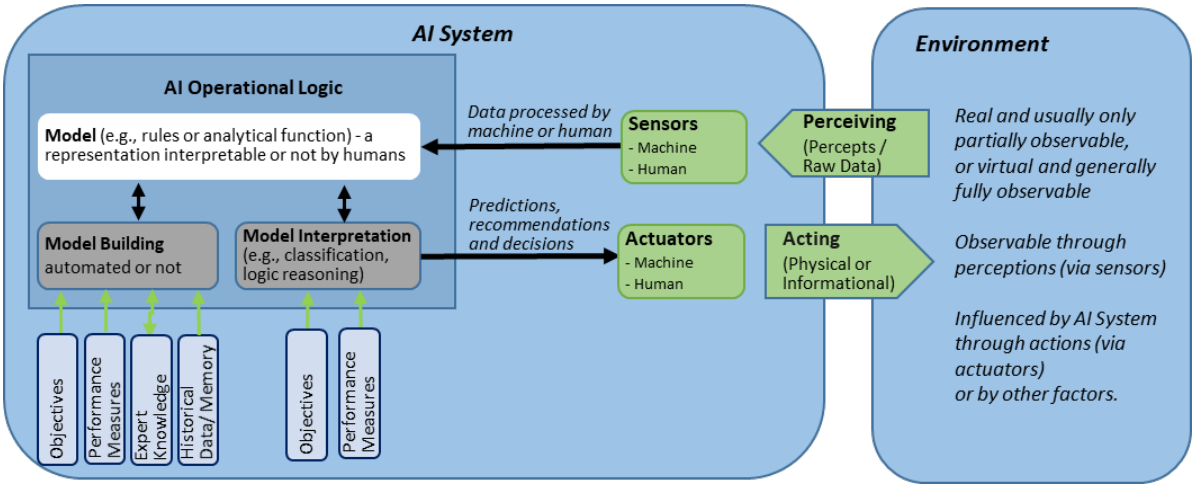
A conceptual view of AI is first presented as the high-level structure of a generic AI system (also referred to as 'Intelligent agent') (Figure 1). An AI system consists of three main elements: Sensors, Operational Logic and Actuators. Sensors collect raw data from the Environment, while Actuators take actions to change the state of the Environment. The key power of an AI system resides in its Operational Logic, which, for a given set of objectives and based on input data from Sensors, provides output for the Actuators – as recommendations, predictions or decisions – that are capable of influencing the state of the Environment.

Figure 1. A high-level conceptual view of an AI system



A more detailed structure captures the main elements that are relevant to the policy dimensions of AI systems (Figure 2). To cover different types of AI systems and different scenarios, the diagram separates the Model Building process (such as machine learning), from the Model (a data object constructed by the Model Building process), and the Model Interpretation process, which uses the Model to make predictions, recommendations and decisions, for the Actuators to influence the Environment.

Figure 2. Detailed conceptual view of an AI System



Environment

An environment in relation to an AI system is a space observable through perceptions (via Sensors) and influenced through actions (via Actuators). Sensors and Actuators are either machines or humans. Environments are either real (*e.g.* physical, social, mental) and usually only partially observable, or virtual (*e.g.* board games) and generally fully observable.

AI system

An AI system is a machine-based system that is capable of influencing the Environment by making recommendations, predictions or decisions for a given set of Objectives. It does so by utilising machine and/or human-based inputs/data to: *i)* perceive real and/or virtual environments; *ii)* abstract such perceptions into models manually or automatically; and *iii)* use Model Interpretations to formulate options for outcomes.

Credit scoring as an illustration of an AI system

A credit-scoring system illustrates a machine-based system that influences its environment (whether people are granted a loan), by making recommendations (a credit score) for a given set of objectives (credit-worthiness). It does so by utilising both machine-based inputs (historical data on people's profiles and on whether they repaid loans) and human-based inputs (a set of rules) to: *i)* perceive real environments (whether people are repaying loans on an ongoing basis); *ii)* abstract such perceptions into models automatically (a credit-scoring algorithm could for example use a statistical model) and *iii)* use model interpretations (the credit-scoring algorithm) to formulate a recommendation (a credit score) of options for outcomes (providing or denying a loan).

"Visually impaired assistant" as an illustration of an AI system

An assistant for visually impaired people illustrates a machine-based system influences its environment by making recommendations (causing a visually impaired person to avoid an obstacle or cross the street) for a given set of objectives (travel from one place to another). It does so utilising machine and/or human-based inputs (large tagged image databases of objects, written words, and even human faces) to: *i)* perceive images of the environment (a camera captures an image of what is in front of a person and sends it to an application), *ii)* abstract such perceptions into models automatically (object recognition algorithms that can recognise a traffic light, a car or an obstacle on the sidewalk) and *iii)* use model interpretation to formulate a recommendation of options for outcomes (providing an audio description of the objects detected in the environment) so the person can decide how to act and thereby influence the environment.

Model

A Model is an actionable representation of all or part of the external environment of an AI system that describes the environment's structure and/or dynamics. The model represents the core of an AI system. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms. Model Interpretation is the process of deriving an outcome from a model.

Model Building

A model can be built or adjusted based on data processed either manually by humans or using automated tools like machine learning algorithms, or both. Model Building often uses Historical Data/Memory to aggregate data automatically into the Model, but can also use Expert Knowledge. Objectives (*e.g.* the output variables) and Performance Measures (*e.g.* accuracy, resources for training, representativeness of the dataset) guide the building process.

Model Interpretation

Model Interpretation is the process by which humans and/or automated tools derive an outcome from the model, in the form of recommendations, predictions or decisions. Objectives and Performance Measures guide the execution. In some cases (*e.g.*, deterministic rules), a model can offers a single recommendation, while in other cases (*e.g.*, probabilistic models), a model can offer a variety of recommendations associated with different levels of, for instance, performance measures like level of confidence, robustness or risk. In some cases, during the interpretation process, it is possible to explain why specific recommendations are made, while in other cases, explanation is almost impossible.

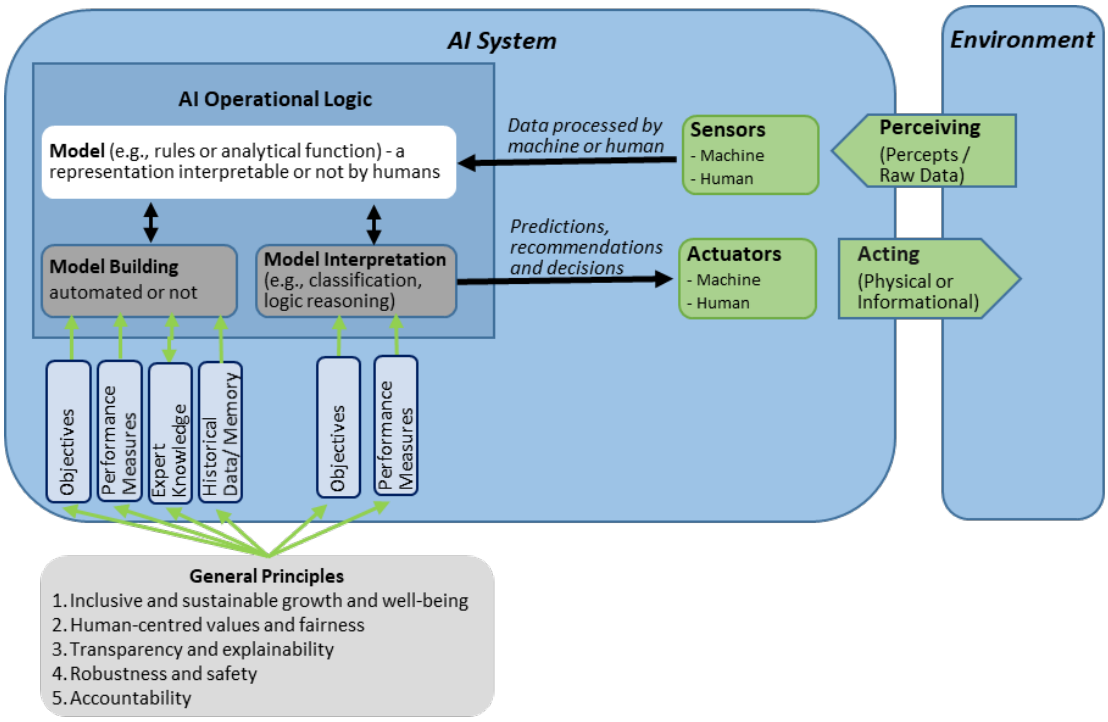
Linking AI Systems to the Principles

The above detailed AI System schema can be linked to the Principles (Figure 3).

Inclusive and sustainable growth and well-being

AI systems can detect patterns in large volumes of data from sensors and can model complex and inter-dependent environments. In turn, AI systems can positively influence the Environment by providing much more accurate and less expensive predictions, recommendations or decisions that generate productivity gains and can help address complex challenges in areas such as science, health and security.

Figure 3. Linking the AI System to the General Principles



Human values and fairness

A model is typically built to achieve specific objectives that may or may not reflect human values, from cancer detection to autonomous weapons. In addition, specific AI systems can be built to achieve a specific set of objectives but later on interpreted with different objectives, as in the case of transfer learning for example.

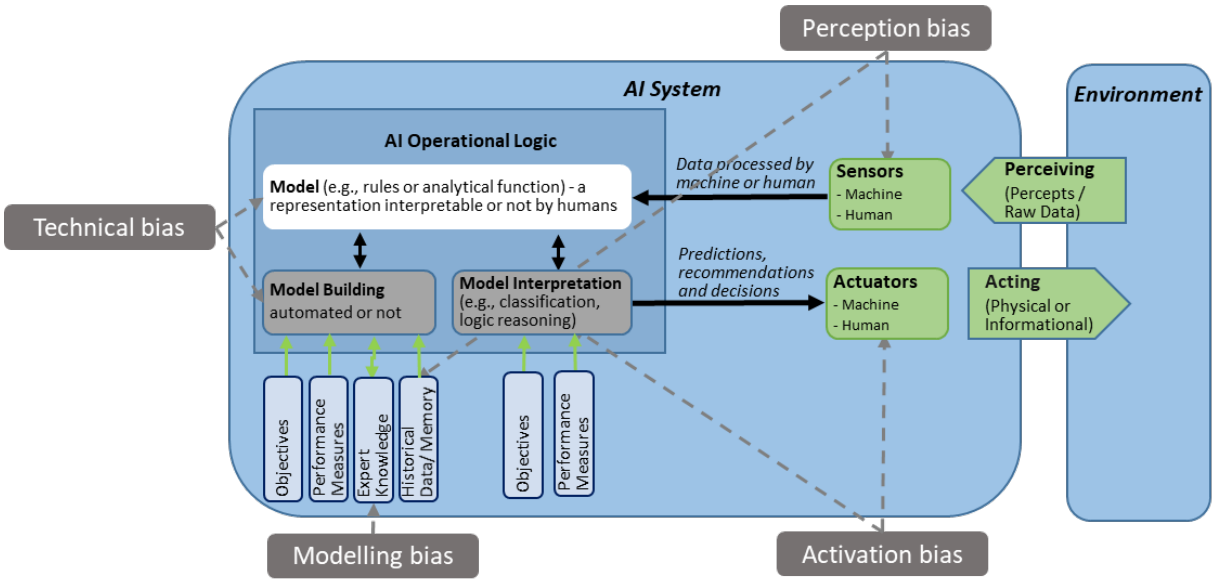
Figure 4 illustrates some of the areas of the AI System in which different types of biases – in particular, perception bias, technical bias, modelling bias and activation bias – are most pronounced. Bias can occur in each of the three main elements of the AI system:

- *Sensors,* notably via Perception bias, whereby the data collected over-represents (or under-represents) one population. Perception bias makes the AI system operate better (or worse) for that population at the expense of others.
- *Operational logic,* notably via Technical bias that arises from constraints or considerations within the technology itself, whether they are known or not. This can include the tools and algorithms an AI system uses. For example, a selected algorithm may work better or worse with a different sets of variables/features. If used in an AI system with different variables or features, its accuracy will be lower, which may introduce bias that is very hard to detect. An accident in 2016 involving a Tesla Model S and a tractor trailer provides an example of technical bias, where the Autopilot's computer vision-based vehicle detection system did not notice the white side of the tractor trailer against a brightly lit sky and did not brake.
- *Expert Knowledge,* notably via Modelling bias, whereby a human manually designing a model (or part of a model) does not take into account some aspects of the environment in building the model, consciously or unconsciously. For example, in an AI system devoted to judiciary decision management, a model can estimate the probability that a person reoffends in future. If the model implemented by a human expert does not take into account

the person's age or gender, for instance because the expert only worked with male or young offenders in the past, the model will include this modelling bias.

- *Actuators*, notably via Activation bias, which relates to how the outputs of the AI system are used in the Environment. For example, actuators such as bots generating twitter posts or news articles can have embedded bias related to the narratives generated by templates.

Figure 4. Areas of the AI system in which biases can appear



Transparency

A Model itself can be interpretable by people (for example in the case of a decision tree) or non-interpretable by people (for example, in the case of deep learning, often referred to as a "black box"). The Model Interpretation process can similarly be more or less understandable. In some cases, during the interpretation process, it is possible to explain why specific recommendations are made, while in other cases (often known as "black box models"), explanation is almost impossible and other types of accountability and transparency measures are called for.

Transparency of an AI system typically focuses on allowing people to understand how an AI system is developed, trained, and deployed; which variables are used, and which variables impact a specific prediction, recommendation or decision.

Robustness and safety

The robustness and safety of AI Systems hinges on Performance Measures that assess how well a system performs compared to specific indicators, for example indicators of accuracy, efficiency, fairness and safety. Performance Measures provide guarantees regarding how a model is built and how it is interpreted. Safety of AI Systems also pertains to Actuators, where most risks of physical and virtual harm reside.

Accountability

Accountability focuses on allocating responsibility to the appropriate organisations or individuals. The accountability of AI systems also relates largely to Performance Measures, which must respect the state of the art.

A Practical Reference Framework for the AI System Lifecycle

In November 2018, the AI Group of experts at the OECD (AIGO) established a subgroup to complement the Principles by detailing the AI system lifecycle. This chapter develops a practical reference framework in which to contextualise and consider ways to implement the Principles in the AI systems lifecycle. After providing an overview of the main phases of the AI system lifecycle, the AI lifecycle actors and the broader set of "stakeholders" affected by AI systems, this annex provides a framework for understanding the risk management approach to AI systems encouraged in the Principles.

Nineteen AI experts participated in the work of the subgroup, which was moderated by Jim Kurose from the U.S. NSF and Nozha Boujemaa from INRIA, and met regularly from mid-December 2018 to mid-February 2019.

The AI System Lifecycle

An AI system incorporates many of the phases involved in traditional software development lifecycles and system development lifecycles more generally but contains specific features.

The AI system lifecycle typically involves the following four phases: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring' (Figure 5).

These phases can be described as follows:

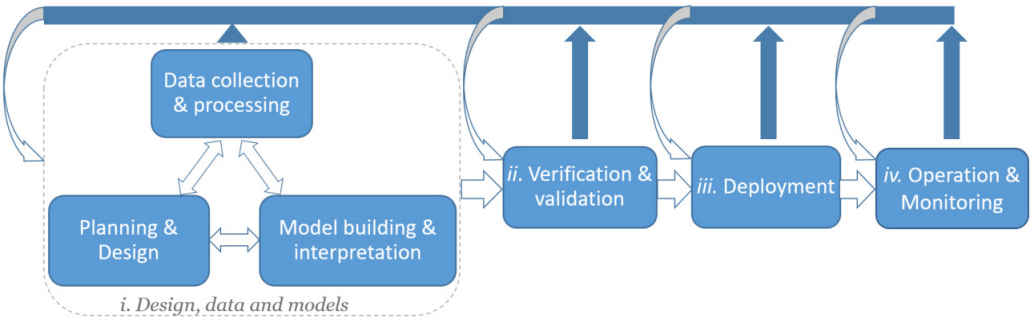
i. **Design, data and modelling** includes several activities, whose order may vary for different AI systems:
   - **Planning and design** of the AI system involves articulating the system's concept and objectives, underlying assumptions, context and requirements, and potentially building a prototype.
   - **Data collection and processing** includes gathering and cleaning data, performing checks for completeness and quality, and documenting the characteristics of the dataset. Dataset characteristics include information on how a dataset was created, its composition, its intended uses, and how it was maintained over time.
   - **Model building and interpretation** involves the creation or selection models/algorithms, their calibration and/or training and interpretation.

ii. **Verification and validation** involves executing and tuning models, with tests to assess performance across various dimensions and considerations.

iii. **Deployment** into live production involves piloting, checking compatibility with legacy systems, ensuring regulatory compliance, managing organisational change, and evaluating user experience.

iv. **Operation and monitoring** of an AI system involves operating the AI system and continuously assessing its recommendations and impacts (both intended and unintended) in light of objectives and ethical considerations. In this phase, problems are identified and adjustments made by reverting to other phases or, if necessary, deciding to retire an AI system from production.

Figure 5. AI system lifecycle



A feature that distinguishes the lifecycle of many AI systems from that of more general system development is the centrality of data and of models that rely on data for their training and evaluation. A characteristic of some AI systems based on machine learning is the capacity to iterate and evolve over time.

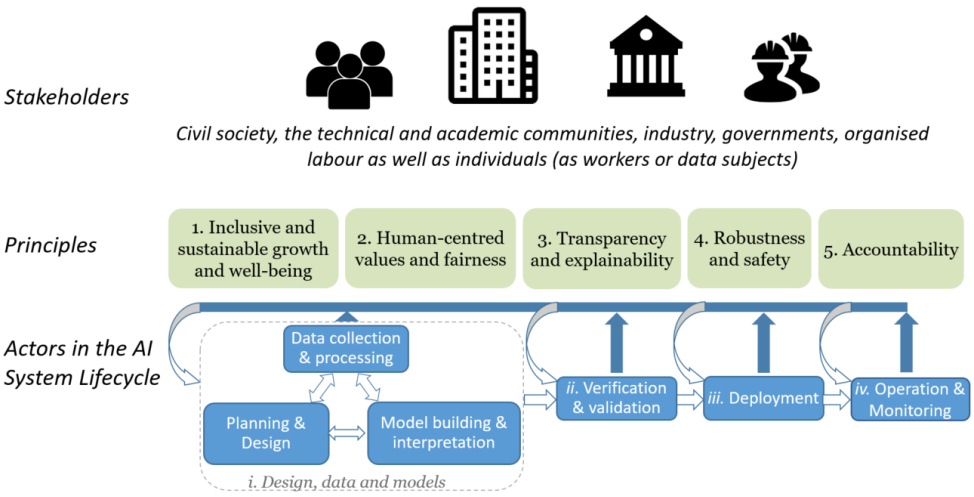
Stakeholders, AI actors and risk management for AI systems

Stakeholders

Stakeholders encompass all public and private sector organisations and individuals involved in, or affected by, AI systems, directly or indirectly. They include, inter alia, civil society, the technical and academic communities, industry, governments, labour representatives and trade unions as well as individuals as workers or data subjects. AI actors are a subset of stakeholders.

Different stakeholders will naturally view each AI principle through a different lens, with different considerations, priorities and questions (Figure 6). These questions and considerations may also differ depending on the phase of the AI system lifecycle.

Figure 6. Stakeholders view of AI principles, in the framework of the AI lifecycle



AI actors

AI actors are those who play an active role in the AI system lifecycle. Public or private sector organisations or individuals that acquire AI systems to deploy or operate them are also considered to be AI actors. AI actors include, inter alia, technology developers, systems integrators, and service and data providers.

The expertise needed at each lifecycle phase varies and may include, inter alia, data science, domain knowledge, modelling, data and model engineering, and governance oversight.

i. **Design, data and modelling:**
   - **Planning and design:** currently involves expertise such as data scientists, domain experts, and governance experts.
   - **Data collection and processing**: currently involves expertise such as data scientists, domain experts, data engineers, data providers.
   - **Model building and interpretation:** currently involves expertise such as modellers, model engineers, data scientists, domain experts.

ii. **Verification and validation:** currently involves expertise such as data scientists, data/model/systems engineers, governance experts.

iii. **Deployment:** currently involves expertise such as system integrators, developers, systems/software engineers and testers.

iv. **Operation and monitoring:** currently involves expertise such as governance experts, domain experts, and systems/software engineers.

A risk management approach for AI systems

Organisations use risk management to identify, assess, prioritise and treat potential risks that can adversely affect the behaviour of systems. Such an approach can also be used to identify risks for different stakeholders and determine how to address these risks throughout the AI system lifecycle.

AI actors implement a risk management approach in conjunction with the AI system lifecycle, both assessing and mitigating risks of the AI system as a whole as well as in each lifecycle phase. As shown in Figure 7, risk management consists of the following steps, whose relevance varies depending on the phase of the AI system lifecycle:

a) *Objectives:* define objectives, functions or properties of the AI system, in context. These functions and properties may change depending on the phase of the AI lifecycle.

b) *Stakeholders and actors:* identify stakeholders and actors involved, i.e., those directly or indirectly affected by the system's functions or properties in each lifecycle phase.

c) *Risk assessment:* assess the potential effects, both benefits and risks, for stakeholders and actors. These will vary, depending on the stakeholders and actors affected, as well as the phase in the AI system lifecycle. In all cases, potential risks to the Principles can be considered.

d) *Risk mitigation:* identify risk mitigation strategies that are appropriate to, and commensurate with, the risk. These should consider factors such as the organisation's goals and objectives, the stakeholders and actors involved, the likelihood of risks materialising and potential benefits.

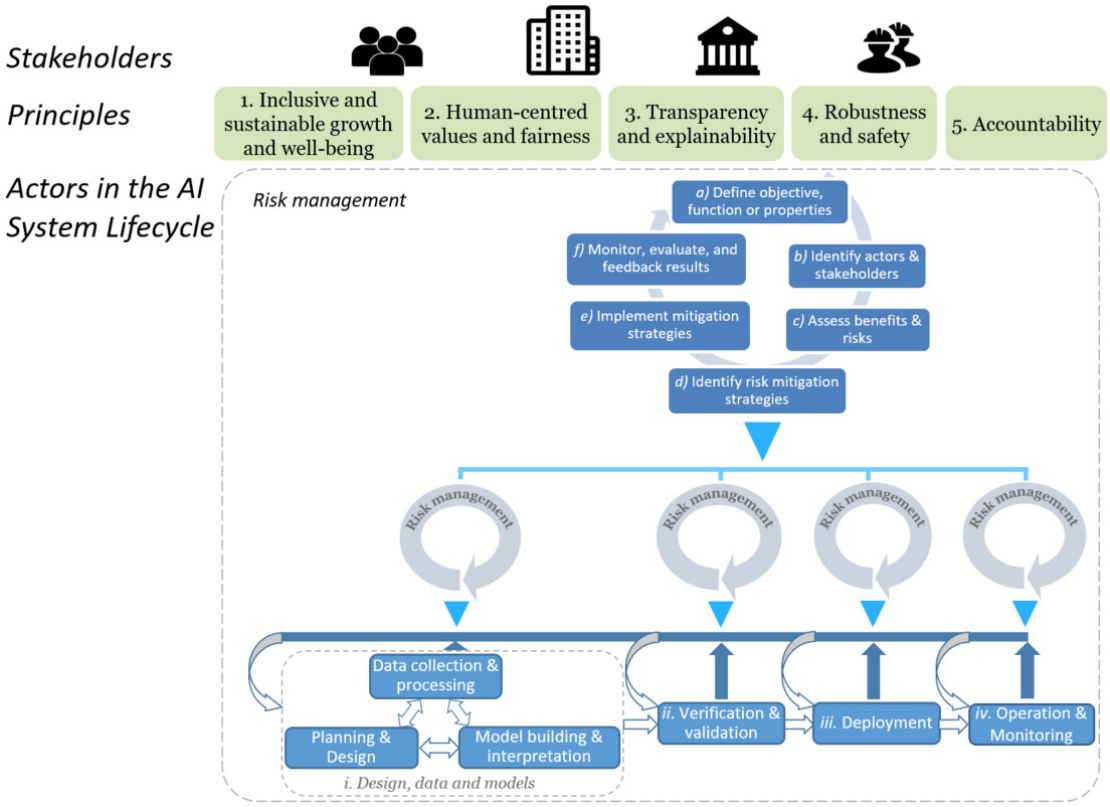
e) *Implementation:* implement risk mitigation strategies.

f) *Monitoring, evaluation and feedback:* monitor, evaluate and feedback results of the implementation.

The use of such an AI risk management system and the documentation of the decisions made at each lifecycle phase can help improve an AI system's transparency and an organisations' accountability for the system.

Figure 7. AI risk-based management approach

To be implemented at each phase of the AI system lifecycle



Illustrating the value of the AI system lifecycle practical reference framework

Fairness considerations provide an example of the value of using such a practical reference framework throughout the AI system lifecycle, to allow stakeholders to engage in more specific discussions and actions in relation to this principle. Different types of biases and other factors that affect fairness may appear in different phases of the AI system lifecycle (Figure 8), including:

i. *Design, data and modelling* phase*:*
   - *Planning and design: non-inclusive design,* whereby an AI system cannot be equally accessed and used by as many people as possible, regardless of age, gender and disability.
   - *Data collection and processing:* data can inaccurately represent the real world or reflect socially-derived artefacts that disadvantage particular groups.
     - *Reporting bias,* whereby people tend to under-report all the information available.
     - *Selection Bias,* whereby the data selected over-represents (or under-represents) one population, making the AI system operate better (or worse) for that population at the expense of others. This can be due, for example, to issues of under-coverage or non-response of some population members or due to the way that sampling is conducted.
     - *Out-group homogeneity bias,* whereby people tend to see those outside their own group as more similar to one another than those in their own groups (e.g. similar attitudes, values, personality traits, and other characteristics).

- *Model engineering, calibration and interpretation:* can also involve biases, notably
  - *Experimenter bias,* whereby the model is subconsciously influenced by the modeller's predisposed notions or beliefs.

ii. *Verification and validation* phase*:*
- *Confirmation bias:* the tendency to search for, interpret, favour, and recall information in a way that confirms one's pre-existing beliefs or hypotheses.

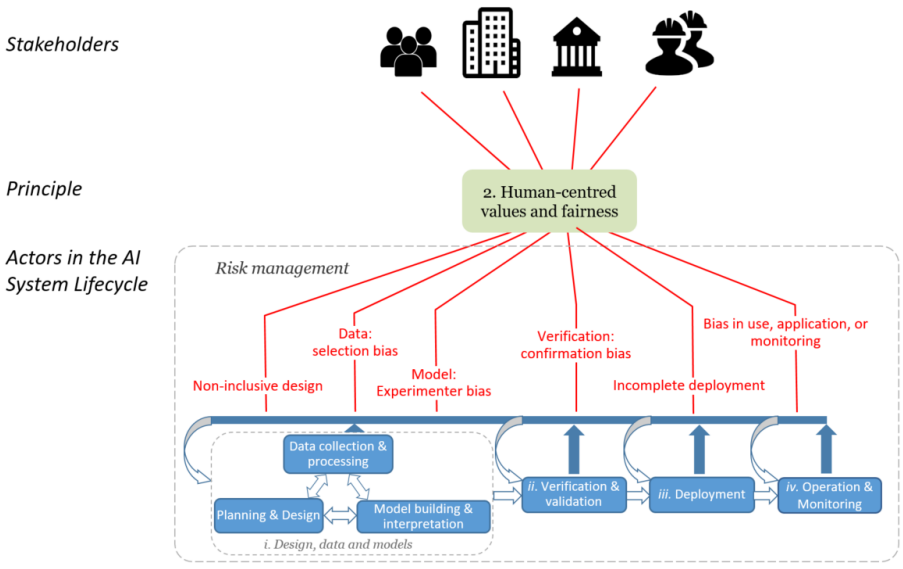
iii. *Deployment* phase*:*
- *Incomplete deployment,* whereby groups of stakeholders may be excluded from using or realising the benefits of a deployed AI system.

iv. *Operation and monitoring* phase*:*
- *Inadequate monitoring,* whereby data may not reflect the breadth of users or uses of a deployed system.

Relevant AI actors can implement risk management strategies to avoid or mitigate these and other biases throughout the AI lifecycle. For example, asking the following questions could help manage and mitigate the risk of *selection bias*:

a) *Defining the objective*: in view of the AI system's objectives, how would selection bias in the data affect its functioning and the Principles?
b) *Stakeholders and actors*: which stakeholders would be affected by selection bias and which AI actors could mitigate this risk?
c) *Risk assessment*: what risks would selection bias create and how likely are they to materialise? What would be the consequences? What level of selection bias would be acceptable in view of potential benefits of the AI system?
d) *Risk mitigation*: what risk controls could be set up during the development phase to prevent selection bias? How can AI actors ensure that this risk remains at an acceptable level?
e) *Implementation*: who should implement the selected risk controls to prevent or mitigate selection bias, when and how?
f) *Monitoring, evaluation and feedback*: how is performance measured, monitored and reviewed, and by who? Who documents and shares information on the risk management of selection bias? With whom?

Figure 8. A view of fairness considerations by AI actors within the AI system lifecycle



Annex A. Scoping principles to foster trust in and adoption of AI

This Annex presents the scoping principles to foster trust in and adoption of artificial intelligence (AI), developed over four meetings by the Expert Group on Artificial Intelligence at the OECD (AIGO). The group concluded its discussion and agreed on this draft at its fourth and last meeting in Dubai, United Arab Emirates, on 8-9 February. This proposal informed the Committee's discussion of a draft Recommendation of the Council on Artificial Intelligence on 14-15 March 2019 that was subsequently adopted by the OECD Council at Ministerial level on 22 May 2019. The present document was declassified by the Committee on 1 July 2019.

Introduction

(References)

Reference to existing OECD instruments (e.g. CDEP + CCP, including privacy and security; MNE Guidelines) and the UN SDGs; UDHR;

Reference to existing national legal, regulatory and policy frameworks applicable to AI, including those related to consumer and personal data protection, intellectual property rights and competition, while noting that such frameworks may need to be adapted;

(Transformative effect of recent developments in AI)

Notably due to recent developments, AI has pervasive, far-reaching and global implications that are transforming societies, economic sectors and the world of work, and are likely to increasingly do so in the future;

(Benefits and challenges)

AI has the potential to improve the welfare of people, to contribute to a positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges, such as climate change, health crises, resource scarcity and discrimination;

At the same time, these transformations may have disparate effects within, and between, societies and economies, notably economic shifts, transitions in the labour market, deepening inequalities, such as gender, income and skills gaps, and detrimental implications on democracy, freedom, fairness, autonomy and individual control, and data privacy and security;

(Need for a global policy framework and practical guidance on AI)

Trust is a key enabler of digital transformation and, while further AI applications and their implications may be hard to foresee, trustworthiness of AI systems is a key factor for diffusion of AI and for capturing the full potential of the technology.

Given the rapid development and implementation of AI, there is a pressing need for a predictable, stable yet adaptive policy environment that promotes a human-centric approach to AI and practical guidance for trustworthy AI, and that applies to all relevant stakeholders according to their responsibility, in a context-sensitive manner.

This policy framework aims to achieve this objective, to empower individuals, public entities, businesses and workers to engage and thereby to create incentives to turn trustworthy AI into a collaborative and competitive parameter in the global marketplace. Striking an appropriate and fair balance between the opportunities offered and the challenges raised by AI applications is essential to steering AI innovation toward inclusive and sustainable growth and well-being, reduction of inequalities between countries and people, and respect of human rights and democratic values.

[**Recognition** that such policy framework should be developed, implemented, monitored and reviewed through continuous international co-operation and multi-stakeholder and interdisciplinary dialogue, that would also guarantee diversity of thought and consideration of national and regional frameworks.]

[**Indication** that:

- the framework below should be regarded as a baseline which can be supplemented by further work from all stakeholders at the OECD and in other fora.

- all principles are inter-related and should be considered as a whole.]

Common understanding of technical terms for the purposes of these principles

AI system

An AI system is a machine-based system that is capable of influencing the environment by making recommendations, predictions or decisions for a given set of objectives.

It does so by utilising machine and/or human-based inputs to: *i)* perceive real and/or virtual environments; *ii)* abstract such perceptions into models manually or automatically; and *iii)* use model interpretations to formulate options for outcomes.

Model

A model is an actionable representation of all or part of the external environment of an AI system that describes the environment's structure and/or dynamics. The model represents the core of an AI system. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms. Model interpretation is the process by which humans and/or automated tools derive an outcome from the model, in the form of recommendations, predictions or decisions.

AI lifecycle

AI system lifecycle phases involve: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'.

AI knowledge

AI knowledge refers to the resources and skills, such as data, algorithms, models, research, know-how, training programmes, governance, processes and best practices, required to understand and participate in the AI system lifecycle.

AI actors

AI actors are those who play an active role in the AI system lifecycle. Public or private sector organisations or individuals that acquire AI systems to deploy, operate and/or use them are also considered to be AI actors.

Stakeholders

Stakeholders encompass all public and private sector organisations and individuals involved in, or affected by, AI systems, directly or indirectly. They include, inter alia, civil society, the technical and academic communities, industry, governments, labour representatives and trade unions as well as individuals as workers or data subjects. AI actors are a subset of stakeholders.

Principles for responsible stewardship of trustworthy AI

1.1. Inclusive and sustainable growth and well-being

All stakeholders should engage in responsible stewardship of trustworthy AI to achieve fair and beneficial outcomes for all people and the planet, such as empowering people and enhancing their capabilities and creativity, advancing inclusion of underrepresented populations and reducing economic and social inequalities, within and across countries, and overall invigorating sustainable economic growth and well-being.

Governments should in particular consider:

- Initiating a meaningful and iterative dialogue inclusive of all stakeholders to enhance understanding of AI, to debate AI-related opportunities and challenges for the economy, the society and the world of work, and to inform policy makers.

- Encouraging AI actors to ensure multidisciplinary collaboration and diversity of views throughout the AI lifecycle to maximise benefits and minimise the potential for harm.

- Supporting AI actors in the implementation of this principle, including through promotion of responsible AI in education and research, exchange of knowledge and best practices, guidance for responsible business conduct and incentives to turn responsible AI into a competitive advantage.

1.2. Human-centred values and fairness

AI actors should set up effective mechanisms to demonstrate respect of human rights and democratic values, including freedom, dignity, autonomy, privacy, non-discrimination, fairness and social justice, and diversity as well as core labour rights, throughout the AI lifecycle.

Governments should in particular consider:

- Encouraging AI actors to assess that AI systems respect human-centred values and fairness on an ongoing basis, and to implement safeguards by design and other measures and processes, including capacity for human final determination, that are appropriate to the context and benefit from multidisciplinary and multi-stakeholder collaboration.

- Promoting codes of ethical conduct, quality standards and quality labels, that help align AI systems with human-centred values and fairness throughout their lifecycle, and help assess AI systems' levels of compliance with these values.

- Ensuring that AI systems allow for individuals' determination over their digital identity and personal data.

1.3. Transparency and explainability

All stakeholders should promote a culture of transparency and responsible disclosure regarding AI systems. In this regard, AI actors should provide, appropriate to the context and state of art, meaningful information to all stakeholders in order to foster understanding of AI systems, to raise their awareness of their interactions with AI systems, including in the workplace, and to enable those adversely affected by an AI system to challenge its recommendations.

Governments should in particular consider:

- Promoting initiatives from AI actors to help make AI systems understandable, including through AI systems that can communicate meaningful information appropriate to the context during their operation to foster understanding of their recommendations.

- Ensuring meaningful disclosure of when and for which purpose stakeholders are interacting with an AI system and who operates it, especially when the system is unbeknownst to the stakeholders.

- Ensuring that natural and legal persons adversely affected by an AI system can obtain, appropriate to the context and state of art, information on the factors and the logic that serve as the basis for its recommendations, without having to comprehend the technology.

1.4. Robustness and safety

AI systems should be robust, in the sense that they should be able to withstand or overcome adverse conditions, and safe, in the sense that they should not pose unreasonable safety risk in normal or foreseeable use or misuse throughout their entire lifecycle.

To this end, AI actors should ensure traceability of the datasets, processes and decisions made during the lifecycle of AI systems to enable understanding of their outcomes and inquiry, where appropriate.

AI actors should also implement or reinforce their risk management approach, on a continuous basis throughout the AI lifecycle, to mitigate risks, including to digital security, as appropriate to the context.

Governments should in particular consider:

- Encouraging AI actors to assess the implications of their contribution to an AI system's lifecycle, in a manner proportionate to their role.

- Calling on AI actors to document the process and decisions made during the AI system's lifecycle, especially for systems with potentially significant consequences on people's lives, to support understanding of AI systems' outcomes and enable accountability.

- Encouraging AI actors to consult stakeholders during the AI lifecycle, including in relation to risk management processes, thus promoting stakeholder participation in all stages of AI systems' lifecycle.

1.5. Accountability

AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their individual role, the context, and state of art.

National policies for trustworthy AI

Governments should develop policies, in co-operation with all stakeholders, to promote trustworthy AI systems and achieve fair and beneficial outcomes for people and the planet, consistent with the principles above.

2.1. Investing in responsible AI research and development

Governments should consider and encourage long-term investments in inter-disciplinary basic research and development to spur innovation in trustworthy AI that would focus on challenging technical issues as well as on AI-related social implications and policy issues.

Governments should in particular consider:

- Developing high-level frameworks to coordinate whole-of-government investments, especially in promising areas underserved by market-driven investments.

- Prioritising inter-disciplinary research and development to address the ethical, legal, and social implications of AI, crosscutting issues such as bias, privacy, transparency, accountability and the safety of AI, and difficult technical challenges such as explainability.

- Building open data sets that are representative and preserve privacy in order to provide an un-biased environment for research and development and to encourage innovation and competition, and opening up existing ones, accordingly.

- Using public procurement, promoting joint public and private procurement, and establishing flexible joint venture funding systems to spur market investment in responsible research and development, to encourage broad-based evolution of the market for AI-based solutions, and to foster diffusion of AI systems that benefit society across regions, firms and demographic groups.

2.2. Fostering an enabling digital ecosystem for AI

Governments should foster an enabling ecosystem, including digital technologies and infrastructure, competitive markets as well as mechanisms for sharing AI knowledge to support the development of trustworthy AI systems.

Governments should in particular consider:

- Investing in, and providing incentives to the private sector to invest in, AI enabling infrastructure and technologies such as high-speed broadband, computing power and data storage, as well as fostering entrepreneurship for trustworthy AI systems.

- Encouraging the sharing of AI knowledge through mechanisms such as open AI platforms and data sharing frameworks while respecting privacy, intellectual property and other rights.

2.3 Providing an agile [and controlled] policy environment for AI

Governments should provide an enabling policy environment to support the agile, safe and transparent transition from research and development to deployment and operation of trustworthy AI systems. To this effect, governments should review existing laws, regulations, policy frameworks and assessment mechanisms as they apply to AI and adapt them, or develop new ones as appropriate.

Governments should further encourage that AI actors comply with the applicable national frameworks and global standards.

Governments should in particular consider:

- Using experimentation, including regulatory sandboxes, innovation centres and policy labs, to provide a controlled environment in which AI systems can be tested.

- Encouraging stakeholders to develop or adapt, through an open and transparent process, codes of conduct, voluntary standards and best practices to guide AI actors throughout the AI lifecycle, including for monitoring, reporting, assessing and addressing harmful effects or misuse of AI systems.

- Establishing and encouraging public and private sector oversight mechanisms of AI systems, as appropriate, such as compliance reviews, audits, conformity assessments and certification schemes, while considering the specific needs of and constraints faced by SMEs.

- Establishing mechanisms for continuous monitoring, reporting, assessing and addressing the implications of AI systems that may pose significant risks or target vulnerable groups.

2.4. Building human capacity and preparing for job transformation

Governments should work closely with social partners, industry, academia, and civil society to prepare for the transition in the world of work and empower people with the competences and skills necessary to use, interact and work with AI.

They should ensure that AI deployment in society goes hand in hand with equipping workers fully for a fair transition and new opportunities in the labour markets. They should do so with a view to fostering

entrepreneurship, creating quality jobs, making human work safer, more productive and more rewarding, and ensuring that no one is left behind.

Governments should in particular consider:

- Developing a policy framework conducive to the creation of new employment opportunities.

- Encouraging research on occupational and organisational changes to anticipate future skills needs and improve safety.

- Promoting a broad, flexible and equal opportunity range of life-long education, technological literacy, skills and capacity-building measures to allow people and workers to successfully engage with AI systems across the breadth of applications.

- Developing schemes, including through social dialogue, for fair transition to support people whose current jobs may be significantly transformed by AI, with a focus on training, career guidance and social safeguard systems.

- Encouraging education institutions and employers to provide interdisciplinary education and training needed for trustworthy AI, from STEM to ethics, including through apprenticeships and reskilling programmes to train AI specialists, researchers, innovators, operators and workers.

2.5 International cooperation for trustworthy AI

Governments should actively cooperate at international level, among themselves and with stakeholders in all countries, to invigorate inclusive and sustainable economic growth and well-being through trustworthy AI in all world regions, and to address global challenges.

They should work together transparently in all relevant global and regional fora to advance the adoption and implementation of these principles and progress on trustworthy AI.

Governments should in particular consider:

- Supporting international and cross-sectoral collaboration concerning these principles, including through open, global multi-stakeholder dialogues that can enable long-term expertise for trustworthy AI.

- Promoting cross-border collaboration for responsible AI innovation through sharing of AI knowledge, and maintaining [free] [transborder] flows of data with trust that safeguard security, privacy, human rights and democratic values.

- Encouraging the development of globally accepted practical technical standards, terminology, taxonomy, and measurement methodologies and indicators to guide international co-operation on trustworthy AI.

- Building AI capacity to bridge digital divides and to share the benefits of trustworthy AI among all countries.

[Provision on measurement to be added: Governments should encourage the development of internationally comparable metrics based on common measurement methodologies, standards and best practices to measure global activity related to AI research, development and deployment, and to gather the necessary evidence base to assess progress in the implementation of these principles.]

Annex B. List of AIGO members

This Annex lists the members and expert contributors to the work of the Expert Group on Artificial Intelligence at the OECD (AIGO).

The following experts contributed to the work of the AIGO as members (Table A B.1). Their contributions are greatly acknowledged.

Table A B.1. AIGO members

Name	Title	Organisation / Country	Group / Delegation
Mr. Wonki Min	[AIGO Chair] Vice-Minister and Chair of the OECD Committee on Digital Economy Policy	Ministry of Science and ICT, Korea	Korea
Mr. Tim Bradley	Minister-Counsellor	Department of Industry, Innovation and Science.	Australia
Mr. Alex Cooke	Counsellor, Department of Industry, Innovation and Science	Australian Embassy to Belgium, Luxembourg and Mission to the European Union and NATO	Australia
Ms. Elissa Strome	Executive Director of the Pan-Canadian AI Strategy	Canadian Institute for Advanced Research (CIFAR)	Canada
Mr. Lars Rugholm Nielsen	Head of Section	Danish Business Authority	Denmark
Mr. Antti Eskola	Commercial Counsellor for Innovation and Enterprise Financing Department	Ministry of Economic Affairs and Employment	Finland
Ms. Christel Fiorina	Head of Audiovisual and Multimedia office	Directorate General for Enterprise, French ministry of economy and finance	France
Mr. Bertrand Pailhes	National Coordinator for the French AI Strategy	State Digital Service, Prime Minister's Service	France
Mr. Michael Schönstein	Head of Strategic Foresight & Analysis	Policy Lab “Digital Work & Society”, Federal Ministry for Labour and Social Affairs	Germany
Mr. Nils Börnsen	Policy adviser responsible for AI policy at BMWI	Federal Ministry for Economic Affairs and Energy	Germany
Mr. András Hlács	Counsellor	Permanent Delegation of Hungary to OECD	Hungary
Mr. Osamu Sudoh	Professor, Graduate School of Interdisciplinary Information Studies	University of Tokyo	Japan
Mr. Susumu Hirano	Dean and Professor	Chuo University Graduate School of Policy Studies	Japan
Mr. Chungwon LEE	Director, Multilateral cooperation division	Ministry of Science and ICT, Korea	Korea
Mr. Seongtak Oh	Executive Director, Department of Bigdata	National Information Society Agency, Korea	Korea
Mr. Javier Juárez Mojica	[Co-moderator, 'What is AI' AIGO subgroup] IFT Commissioner	Federal Telecommunications Institute	Mexico
Mr. Wim Rullens	Senior Policy Coordinator	Ministry of Economic Affairs and Climate	Netherlands
Ms. Olivia Erdelyi	Lecturer	Canterbury University	New Zealand
Mr. Robert Kroplewski	Representative	Minister for Digitalisation of the Information Society in Poland	Poland
Mr. Andrey Ignatyev	Deputy Head of OECD Unit	Ministry of Economic Development	Russian Federation
Mr. Konstantin Vishnevskiy	Head of Department for Digital Economy Studies ISSEK HSE	Institute for Statistical Studies and Economics of Knowledge	Russian Federation
Mr. Yeong Zee Kin	Assistant Chief Executive (Data Innovation and Protection Group)	Infocomm Media Development Authority (IM;DA), Government of Singapore	Singapore
Mr. Michal Cíž	AI Policy Expert, EU Digital Single Market	Deputy Prime Minister’s Office for Investments and Informatization	Slovak Republic
Mr. Marko Grobelnik	[Co-moderator, 'What is AI' AIGO subgroup] Researcher in AI	Jozef Stefan Institute - Artificial Intelligence Lab	Slovenia

Ms. Helena Hånell McKelvey	Head of Section, Division for Digital Development	Ministry of Enterprise and Innovation	Sweden
Ms. Livia Walpen	Advisor, International Relations	Swiss Federal Office of Communications	Switzerland
Ms. Ezgi Bener	Expert on Scientific Programmes	The Scientific and Technological Research Council of Turkey (TUBITAK)	Turkey
Mr. Cyrus Hodes	Advisor to the UAE Minister for AI	UAE Ministry for AI	United Arab Emirates
Mr. Edward Teather	Senior Policy Adviser	Office for Artificial Intelligence	United Kingdom
Mr. Adam Murray	International Affairs Officer, Office of International Communications and Information Policy	U.S. Department of State	United States
Ms. Fiona Alexander	NTIA Associate Administrator	U.S. Department of Commerce	United States
Mr. Jim Kurose	[Co-moderator, 'AI system lifecycle' AIGO subgroup] Assistant Director for Computer and Information Science and Engineering, Assistant Director for AI at the Office of Science and Technology Policy	U.S. National Science Foundation	United States
Mr. Matt Chessen	A/Deputy Science and Technology Adviser to the Secretary of State	U.S. Department of State	United States
Ms. Irina Orssich	Political Analyst	European Commission	European Commission
Mr. Jean-Yves Roger	Policy Officer	European Commission	European Commission
Mr. Barry O'Brien	Government and Regulatory Affairs Executive	IBM (Ireland)	BIAC
Ms. Carolyn Nguyen	Director, Technology Policy Group	Microsoft	BIAC
Mr. Ludovic Peran	Public Policy & Gov't Relations	Google	BIAC
Mr. Noberto Andrade	Privacy and Public Policy Manager	Facebook	BIAC
Mr. Marc Rotenberg	Executive Director	Electronic Privacy Information Center (EPIC)	CSISAC
Mr. Suso Baleato	Secretary	CSISAC	CSISAC
Mr. Konstantinos Karachalios	Managing Director	IEEE	ITAC
Ms. Anna Byhovskaya	Senior Policy Advisor	TUAC - Trade Union Advisory Committee to the OECD	TUAC
Ms. Christina J. Colclough	Director Platform & Agency Workers, Digitalisation and Trade	Uni Global Union (UNI)	TUAC
Mr. Nicolas Miailhe	Co-Founder of AI Initiative	AI Initiative (civil society)	Invited expert
Ms. Verity Harding	Co-Lead	DeepMind Ethics & Society	Invited expert
Mr Jason Stanley	Design Research Practice Lead	ElementAI	Invited expert
Mr. Urs Gasser	Director, Technology Policy Group	Harvard Berkman Klein Center	Invited expert
Mr. Ryan Budish	Senior Researcher	Harvard Berkman Klein Center	Invited expert
Ms. Nozha Boujemaa	[Co-moderator, 'AI system lifecycle' AIGO subgroup] Director of Research	INRIA	Invited expert
Mr. Michel Morvan	President / Executive Chairman	IRT SystemX / Cosmo Tech	Invited expert
Mr. Taylor Reynolds	Director, Technology Policy	MIT	Invited expert
Mr. Danny Weitzner	Principal Research Scientist	MIT	Invited expert
Mr. Jonathan	PhD Candidate	MIT	Invited expert

Frankle			
Mr. Jack Clark	Policy Director	OpenAI	Invited expert
Mr Dudu Mimran	CTO	Telekom Innovation Laboratories Israel	Invited expert
Mr. Moez Chakchouk	Assistant Director-General for Communication and Information	UNESCO	Invited expert
Ms. Pam Dixon	Founder/ executive director	World Privacy Forum	Invited expert

The work of AIGO benefited from the contributions and input of other experts (Table A B.2). We gratefully acknowledge their contributions.

Table A B.2. Other contributors to AIGO

Name	Title	Organisation / Country	Group / Delegation
Ms. Karen McCabe	Senior Director, Technology Policy and International Affairs	IEEE	ITAC
Mr. Kentaro Kotsuki	Director of the Policy Research Department Institute for Information and Communications Policy (IICP)	Ministry of Internal Affairs and Communications	Japan
Mr. Tomáš Jucha	Director of Department of Innovative Technologies and International Cooperation	Deputy Prime Minister's Office for Investments and Informatization of the Slovak Republic	Slovak Republic
Mr. Timotej Šooš	Key Horizontal Projects Coordinator	Ministry of Foreign Affairs of Slovenia	Slovenia
Mr. Daniel Egloff	Professor	University of Lausanne	Switzerland
Mr. Philippe Labouchère	Project Leader for Innovation & Entrepreneurship	Swissnex Boston	Switzerland
Mr. Kelly Ross	Deputy Policy Director	American Federation of Labor and Congress	TUAC
Mr. Doug Franz		IEEE	Invited expert
Ms. Eva Thelisson	Co-Founder & CEO	AI Transparency Institute	Invited expert

In addition, we thank the following experts for their contributions to the work of the AIGO subgroups:

Name	Title	Organisation / Country
Mr. Wael Diab	Chair SC 42 (Artificial Intelligence)	ISO
Mr. James Hodson	Member of the Board of Directors and CEO	AI for Good foundation
Mr. Ali G Hessami	Chair and Tech Editor, IEEE P7000 Tech-Ethics Standard	IEEE
Mr. Abe Hsuan	IT & IP lawyer	
Mr. Grigory Marshalko	Expert of the Technical committee for standardization "Cryptography and security mechanisms", "IT security techniques", and "AI"	ISO
Mr. John Shawe Taylor	Head of Computer Science department at UCL and UNESCO AI Chair	UCL (University College London)
Ms. Ingrid Volkmer	Professor and Head, Media and Communications Program	University of Melbourne
Mr. Michael Witbrock	Head, AI Foundations Lab - Reasoning	IBM Research AI

The support of MIT Internet Policy Research Initiative and of the UAE Ministry for AI, which each hosted an AIGO meeting, is also gratefully acknowledged.

[1] The meeting at MIT in January 2019 was chaired by Ms. Fiona Alexander.