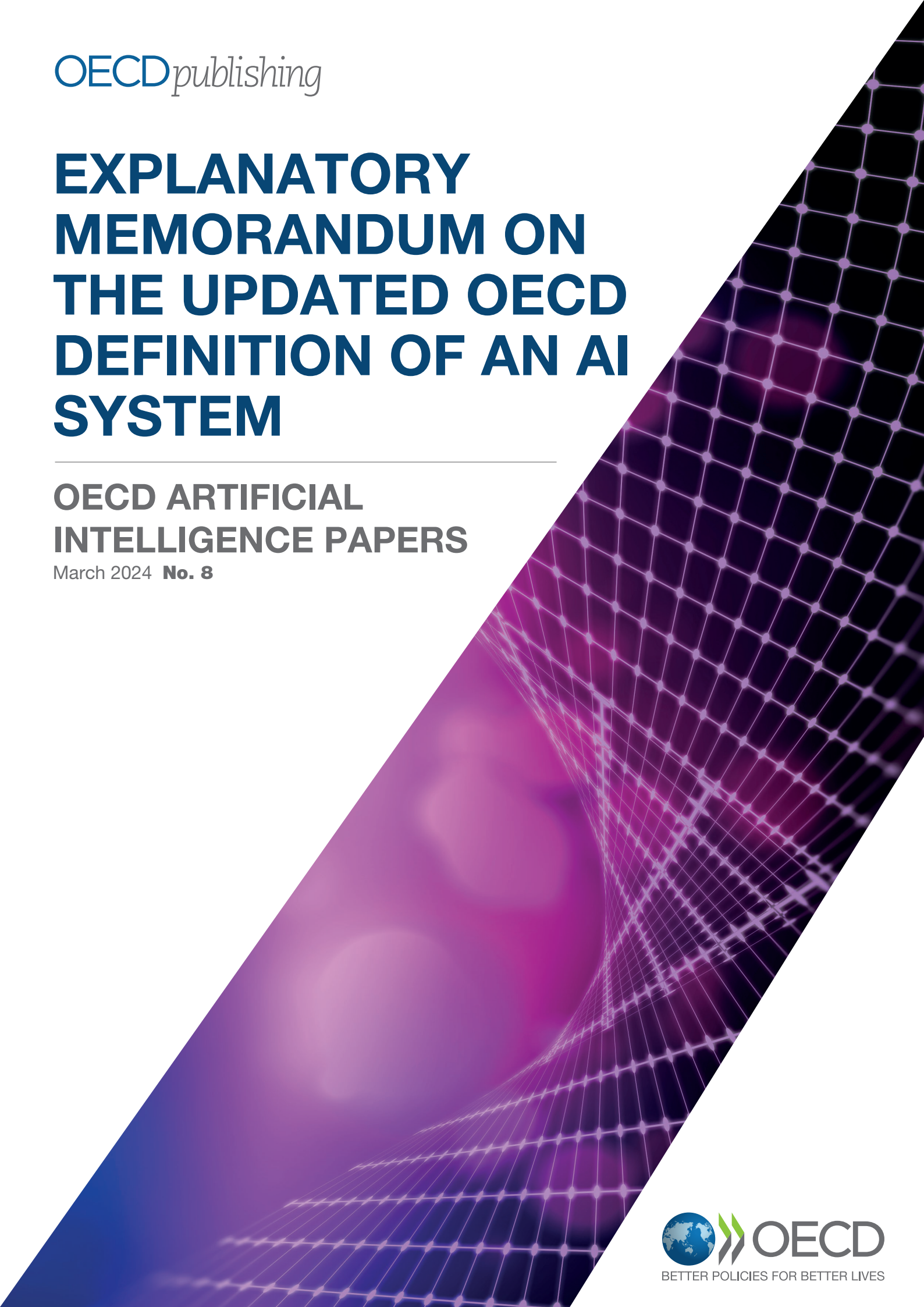
OECD publishing

EXPLANATORY MEMORANDUM ON THE UPDATED OECD DEFINITION OF AN AI SYSTEM

OECD ARTIFICIAL INTELLIGENCE PAPERS
March 2024 **No. 8**



OECD
BETTER POLICIES FOR BETTER LIVES

This paper was approved and declassified by written procedure by the Committee on Digital Economy Policy on 15 December 2023 upon the proposal of the OECD Working Party on AI Governance (AIGO), and prepared for publication by the OECD Secretariat.

Note to Delegations:

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

DSTI/CDEP/AIGO(2023)8/FINAL

ACKNOWLEDGEMENTS

The *OECD Explanatory Memorandum on the Updated OECD Definition of an AI System* complements the revised definition of an AI system and provides further technical background. While the definition is necessarily short and concise, its application in practice depends on a range of complex and technical considerations. The Explanatory Memorandum aims to support all adherents to the OECD AI Principles for better implementation.

The report was prepared by Karine Perset, Head of the Artificial Intelligence Unit in the OECD Digital Economy Policy Division, with support from Lucia Russo, Luis Aranda, Yuki Yokomori and Gallia Daor, and under the guidance of Audrey Plonk, Deputy Director of the OECD Science, Technology and Innovation Directorate.

The AIGO meetings leading up to the agreement on the updated definition of an AI system and on the explanatory memorandum were chaired by Marco-Alexander Breit from Germany and Juraj Čorba from Slovakia. The report benefitted greatly from the inputs of Marko Grobelnik, expert researcher at the AI Lab of Slovenia's Jožef Stefan Institute (JSI) and from Stuart Russel, Professor of Computer Science, University of California, Berkeley. The report also benefitted greatly from the inputs of the delegations participating in the Working Party on AI Governance. Among many, the drafting team wishes to thank Jesse Dunietz, Tatjana Evans, Emilia Gomez, Samo Zork, Elham Tabassi, Mark Latonero, David Turnbull, Nobuhisa Nishigata, Robert Kroplewski, Katinka Clausdatter Worsøe, Zumrut Muftuoglo, Arturo Robles Rovalo, Amit Thapar, Sebastian Hallensleben, Pam Dixon, and Barry O'Brien.

The OECD is grateful to John Tarver (consultant to the OECD) for editing this report, and to Andreia Furtado for editorial and publishing support. The report benefited significantly from their engagement.

1 BACKGROUND AND UPDATED DEFINITION OF AN AI SYSTEM IN THE OECD RECOMMENDATION ON AI

This document contains clarifications to the definition of an AI system contained in the 2019 OECD Recommendation on AI (the "AI Principles") (OECD, 2019[1]) to support their continued relevance and technical soundness. On 9 November 2022, 20 April 2023, 2 July 2023, 11 September 2023 and 6 October 2023, AIGO delegates and other technical experts discussed clarifications to the definition in dedicated sessions and workshops. At a joint Session of the Working Party on AI Governance and the Committee on Digital Economy Policy of 16 October, the Committee agreed to transmit the revised definition to the OECD Council for adoption on 8 November 2023.

These updates are the first of two steps to review the implementation, dissemination, and continued relevance of the Recommendation, required by the Council at the five-year mark of their adoption (2024). The goal of the definition of an AI system in the OECD Recommendation is to articulate what is considered to be an AI system for the purposes of the recommendation. The review and revision of the definition was undertaken earlier than the rest of the review due to its fundamental importance to the review process. Earlier consideration also endeavoured to support broad alignment of the definitions of AI systems in several ongoing processes in the European Union, Japan, and other jurisdictions and organisations that are seeking to improve governance of AI. The second part of the review, *i.e.*, the full report on the implementation, dissemination, and continued relevance of the Recommendation, was discussed by AIGO and CDEP starting in Fall 2023.

The OECD definition of an AI system contained in the OECD AI Principles (OECD, 2019[1]); (OECD, 2019[2]) built on the conceptual view of AI detailed in Artificial Intelligence: A Modern Approach (Russell and Norvig, 2009[3]). It read: "*An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy*".

The text above is replaced with the following updated definition:

> *An AI system is a machine-based system that ~~can~~, for ~~a given set of human-defined~~ explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as ~~makes~~ predictions, content, recommendations, or decisions that can influenc~~ing~~e physical ~~real~~ or virtual environments. Different AI systems ~~are designed to operate with~~ vary~~ing~~ in their levels of autonomy and adaptiveness after deployment.*

The updated definition reads as follows:

> ***An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.***

In addition to updating the definition slightly, the present Explanatory Memorandum on the definition of an AI system contains information to help to interpret the definition by adding more information on how AI systems are built and operate. It builds on expert input, on the OECD Classification Framework (OECD, 2022[4]), and on the work of the OECD expert group on "what is AI" undertaken in 2018-2019 (OECD, 2019[5]). This explanatory memorandum will accompanies the Recommendation but is not part of it and is not be submitted to the Council for adoption.

2 EXPLANATORY MEMORANDUM

> ***An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.***

TOPICS TYPICALLY ENCOMPASSED BY THE TERM "AI"

Topics typically encompassed by the term “AI” and in the definition of an AI system include categories of techniques such as machine learning and knowledge-based approaches, and application areas such as computer vision, natural language processing, speech recognition, intelligent decision support systems, intelligent robotic systems, as well as the novel application of these tools to various domains. AI technologies are developing at a rapid pace and additional techniques and applications will likely emerge in the future. The OECD definition aims to be flexible by reflecting a broad understanding of AI, and actors using this definition are encouraged to exercise judgement on its relevant scope, depending on the context it is being used in.

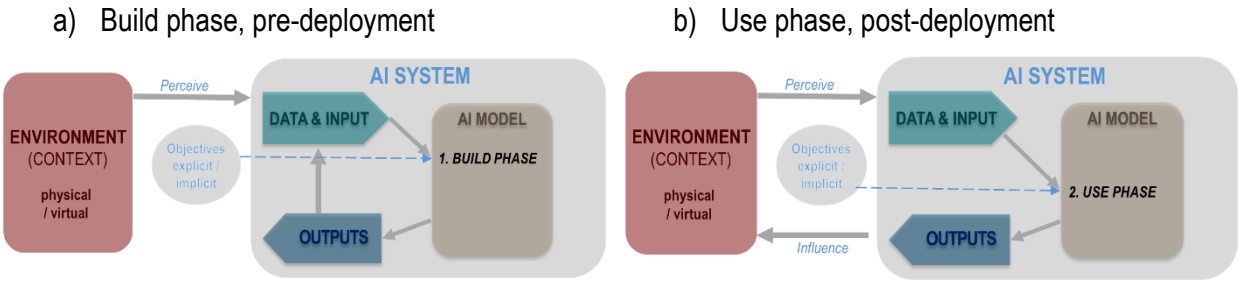
ROLE OF HUMANS, AUTONOMY AND ADAPTIVENESS

Regarding the role of humans, an AI system’s objective setting and development can always be traced back to a human who originates the AI system development process, even when the objectives are implicit. However, some types of AI systems can develop implicit sub-objectives and sometimes set objectives for other systems. Human agency, autonomy, and oversight vis-à-vis AI systems are critical values in the OECD AI Principles that depend on the context of AI use. The OECD definition of an AI system intentionally does not address the issue of liability and responsibility for AI systems and their potentially harmful effects, which ultimately rests with humans and does not in any way pre-determine or pre-empt regulatory choices made by individual jurisdictions in that regard.

AI system autonomy (contained in both the original and the revised definition of an AI system) means the degree to which a system can learn or act without human involvement following the delegation of autonomy and process automation by humans. Human supervision can occur at any stage of the AI system lifecycle, such as during AI system design, data collection and processing, development, verification, validation, deployment, or operation and monitoring. Some AI systems can generate outputs without these outputs being explicitly described in the AI system’s objective and without specific instructions from a human.

Adaptiveness (contained in the revised definition of an AI system) is usually related to AI systems based on machine learning that can continue to evolve after initial development. The system modifies its behaviour through direct interaction with input and data before or after deployment. Examples include a speech recognition system that adapts to an individual’s voice or a personalised music recommender system. AI systems can be trained once, periodically, or continually and operate by inferring patterns and relationships in data. Through such training, some AI systems may develop the ability to perform new forms of inference not initially envisioned by their programmers.

Figure 1. Illustrative, simplified overview of an AI system



Source: (OECD, 2022[6])

Note: this figure presents only one possible relationship between the development and deployment phases. In many cases, the design and training of the system may continue in downstream uses. For example, deployers of AI systems may fine-tune or continuously train models during operation, which can significantly impact the system's performance and behaviour.

ENVIRONMENT OR CONTEXT

An environment or context in relation to an AI system is an observable or partially observable space perceived using data and sensor inputs and influenced through actions (through actuators). The environments influenced by AI systems can be physical or virtual and include environments describing aspects of human activity, such as biological signals or human behaviour. Sensors and actuators are either humans or components of machines or devices.

AI SYSTEM OBJECTIVES

AI system objectives can be explicit or implicit[1]; for example, they can belong to the following categories, which may overlap in some systems:

- *Explicit and human-defined* – where the developer encodes the objective directly into the system (*e.g.*, through an objective function). Examples of systems with explicit objectives include simple classifiers, game-playing systems, reinforcement learning systems, combinatorial problem-solving systems, planning algorithms, and dynamic programming algorithms.
- *Implicit in (typically human-specified) rules* – rules dictate the action to be taken by the AI system according to the current circumstance. For example, a driving system might have a rule, "If the traffic light is red, stop." However, these systems' underlying objectives, such as compliance with the law or avoiding accidents, are not explicit, even though they are typically human-specified.
- *Implicit in training data* – where the ultimate objective is not explicitly programmed but incorporated through training data and a system architecture that learns to emulate those data (*e.g.*, rewarding large language models for generating a plausible response)[2];
- *Not fully known in advance* – some examples include recommender systems that use reinforcement learning to gradually narrow down a model of individual users' preferences.

AI systems can operate according to one or more types of objectives. In addition, user prompts can supplement design objectives when the system is in operation (*e.g.*, a specific destination supplied by the user in the case of a navigation system or prompts in large language models).

It can be challenging to specify explicit objectives for AI systems that implement the system's intent, *i.e.*, the explicit objective may result in unanticipated consequences. It is often the case that spelling out the

designer's true objectives can be too difficult or result in a system that is inefficient. Thus, the objectives used to create the systems are often higher-level objectives. The degree to which the explicit objective and the designers' intent differ is sometimes referred to as misalignment.

INPUT, INCLUDING DATA

Input is used both during development and after deployment. Input can take the form of knowledge, rules and code humans put into the system during development or data. Humans and machines can provide input.

During development, input is leveraged to build AI systems, *e.g.*, with machine learning that produces a model from training data and/or human input. Input is also used by a system in operation, for instance, to infer how to generate outputs. Input can include data relevant to the task to be performed or take the form of, for example, a user prompt or a search query. [3]

BUILDING AI SYSTEMS AND MODELS

Prior to deployment, an AI system is typically built by combining one or more models developed manually or automatically (*e.g.*, with reasoning and decision-making algorithms) based on machine and/or human inputs/data.

- Machine learning is a set of techniques that allows machines to improve their performance and usually generate models in an automated manner through exposure to training data, which can help identify patterns and regularities rather than through explicit instructions from a human. The process of improving a system's performance using machine learning techniques is known as "training".
- Symbolic or knowledge-based AI systems typically use logical and/or probabilistic representations, which may be human-generated or machine-generated, to infer outputs such as predictions and decisions. These representations rely on explicit descriptions of variables and of their interrelations. For example, a system that reasons about manufacturing processes might have variables representing factories, goods, workers, vehicles, machines, and so on.
- In addition, symbolic AI may use machine learning. For example, inductive logic programming learns symbolic logical representations from data, and decision-tree learning learns symbolic logical rules in the form of a tree of logical conditions.

Although different interpretations of the word "model" exist, in this document, an AI model is a core component of an AI system used to make inferences from inputs to produce outputs. It is important to note that while the parameters of an AI model change during the build phase, they usually remain fixed after deployment once the build phase has concluded. However, some AI systems composed of model(s) built using machine learning methods can continue to adapt after the initial build phase, improving their performance by interacting directly with new input and data. Furthermore, AI systems may be periodically updated/retrained, re-tested, and re-deployed as new versions.

A model is defined as a "physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data" in the ISO/IEC 22989 standard. AI models include, among others, statistical models and various kinds of input-output functions (such as decision trees and neural networks). An AI model can represent the transition dynamics of the environment, allowing an AI system to select actions by examining their possible consequences using the model.[4] AI models can be built manually by human programmers or automatically through, for example, unsupervised, supervised, or reinforcement machine learning techniques.

In addition, although AI systems are constructed using machine learning methods or logic- and knowledge-based approaches, they generally do not have access to any training data after deployment.

"INFERRING HOW TO" GENERATE OUTPUTS

The concept of "inference" generally refers to the step in which a system generates an output from its inputs, typically after deployment.[5] When performed during the build phase, inference, in this sense, is often used to evaluate a version of a model, particularly in the machine learning context. In the context of this explanatory memorandum, "infer how to generate outputs" should be understood as also referring to the build phase of the AI system, in which a model is derived from inputs/data.

OUTPUT(S)

The output(s) generated by an AI system generally reflect different functions performed by AI systems. AI system outputs generally belong to the broad categories of recommendations, predictions, and decisions. These categories correspond to different levels of human involvement, with "decisions" being the most autonomous type of output (the AI system affects its environment directly or directs another entity to do so) and "predictions"[6] the least autonomous. For example, a driver-assist system might "predict" that a pixel region in its camera input is a pedestrian; it might "recommend" braking, or it might "decide" to apply the brake. Generative AI systems that produce "content" —including text, image, audio, and video—have gained significant momentum. Although one could, for example, view the generation of text as a sequence of decisions to output particular words (or predictions of words that would be likely to appear in a specific context), content generation systems have become such an important class of AI systems that they are included as their own output category in the present revised definition.

Box 1. Different types of tasks performed by AI systems

Outputs generally reflect different tasks or functions performed by AI systems. They include, but are not limited to, recognition (identifying and categorising data, *e.g.*, image, video, audio and text, into specific classifications as well as image segmentation and object detection), event detection (connecting data points to detect patterns, as well as outliers or anomalies), forecasting (using past and existing behaviours to predict future outcomes), personalisation (developing a profile of an individual and learning and adapting its output to that individual over time), interaction support (interpreting and creating content to power conversational and other interactions between machines and humans, possibly involving multiple media such as voice text and images), goal-driven optimisation (finding the optimal solution to a problem for a cost function or predefined goal) and reasoning with knowledge structures (inferring new outcomes that are possible even if they are not present in existing data, through modelling and simulation).

Source: (OECD, 2022[4])

APPLICATION OF THE UPDATED DEFINITION

The updated definition of AI is inclusive and encompasses systems ranging from simple to complex. AI represents a set of technologies and techniques applicable to many different situations. Specific techniques, such as machine learning, may raise particular considerations for policymakers, such as bias, transparency, and explainability, and some contexts of use (*e.g.*, decisions about public benefits) may raise more significant concerns than others. For that reason, when applied in practice, additional criteria may be needed to narrow or otherwise tailor the definition when used in a specific context.

References

OECD (2022), *OECD Framework for the Classification of AI systems*, OECD Publishing, https://doi.org/10.1787/cb6d9eca-en. [4]

OECD (2022), "OECD Framework for the Classification of AI systems", *OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en. [6]

OECD (2019), *Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449*, https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449. [1]

OECD (2019), *Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)*, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en. [5]

OECD (2019), "Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)", *OECD Digital Economy Papers*, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en. [2]

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Penguin Publishing Group, 2019. [7]

Russell, S. and P. Norvig (2009), *Artificial Intelligence: A Modern Approach, 3rd edition*, Pearson, London, http://aima.cs.berkeley.edu/. [3]

End notes

[1] AI systems are supplied with objectives. However, the ultimate human objectives (e.g., respecting traffic rules) may be higher-level or more abstract than the technical objectives supplied to the system (e.g., stopping at a red light), which should serve the higher-level objectives.

[2] From a more technical perspective, the learning process for these imitative systems relies on a formally defined loss function that measures the deviation of the system's outputs from observed human behaviour. The loss function does not, however, express the human objectives underlying the observed behaviour. This is in contrast with a reinforcement learning approach, where the reward function used for training may explicitly express the humans' ultimate objectives and trade-offs.

[3] Some experts have pointed out that one type of AI system malfunction related to its input may involve a system partially or totally ignoring its input data when generating its output.

[4] With respect to this sense of the word "model", the phrase "model-free" (as in "model-free reinforcement learning") refers to a system that operates without a transition model for its environment but may instead use a direct policy mapping from inputs to actions, a Q-function mapping from inputs and actions to values, or some other way to choose actions. In such cases, the policy or Q-function would be an "AI model" in the more general sense of an input-output function.

[5] According to ISO 22989, "inference" refers both to the process of reasoning to derive conclusions from known premises (facts, rules, models, features or raw data) and to the result of that process.

[6] In the context of AI, "prediction" does not necessarily mean making a guess about the future. The term can simply mean making a guess about an unknown value (the output) from known values supplied to the system (the input).