The OECD Framework for the Classification of AI systems

The OECD Framework for the Classification of AI Systems helps characterise AI systems to identifiy policy opportunities and challenges

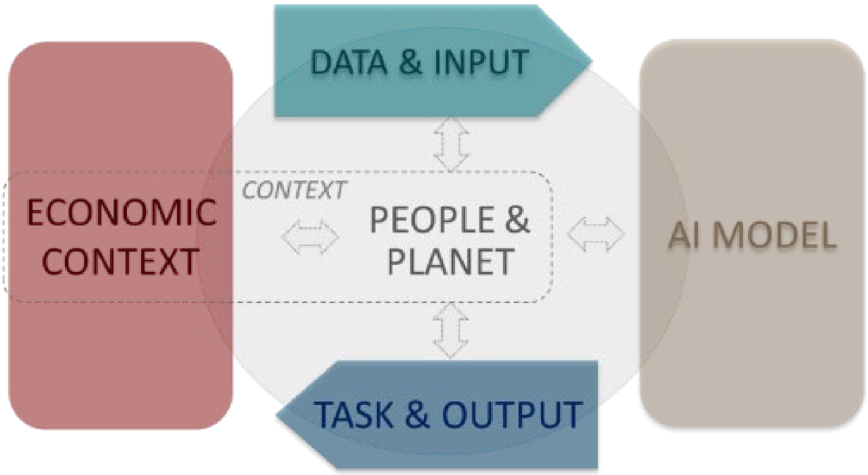
AI changes how people learn, work, play, interact and live. As AI spreads across sectors, different types of AI systems deliver different benefits, risks and policy and regulatory challenges. Consider the differences between a virtual assistant, a self-driving vehicle and an algorithm that recommends videos for children.

The OECD developed a user-friendly framework for policy makers, regulators, legislators and others to characterise AI systems for specific projects and contexts. The framework links AI system characteristics with the OECD AI Principles (OECD, 2019), the first set of AI standards that governments pledged to incorporate into policy making and promote the innovative and trustworthy use of AI.

Key dimensions structure AI system characteristics and interactions

The framework classifies AI systems and applications along the following dimensions: People & Planet, Economic Context, Data & Input, AI Model and Task & Output. Each one has its own properties and attributes or sub-dimensions relevant to assessing policy considerations of particular AI systems.

Key dimensions of the OECD Framework for the Classification of AI Systems



Ways to use the framework

The framework allows users to zoom in on specific risks that are typical of AI, such as bias, explainability and robustness, yet it is generic in nature. It facilitates nuanced and precise policy debate. The framework can also help develop policies and regulations, since AI system characteristics influence the technical and procedural measures they need for implementation. In particular, the framework provides a baseline framework to help support and advance:

- A common understanding of AI, and metrics
- Registries or inventories of AI systems
- Sector-specific frameworks, e.g. in healthcare
- Risk assessment, incident reporting and risk management

Classification framework dimensions and criteria at a glance

PEOPLE & PLANET	Criteria	Description
USERS	Users of AI system	What is the level of competency of users who interact with the system?
STAKEHOLDERS	Impacted stakeholders	Who is impacted by the system (e.g. consumers, workers, government agencies)?
OPTIONALITY	Optionality and redress	Can users opt out, e.g. switch systems? Can users challenge or correct the output?
HUMAN RIGHTS	Human rights and democratic values	Can the system's outputs impact fundamental human rights (e.g. human dignity, privacy, freedom of expression, non-discrimination, fair trial, remedy, safety)?
WELL-BEING & ENVIRONMENT	Well-being, society and the environment	Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?
DISPLACEMENT	*{Displacement potential}*	*Could the system automate tasks that are or were being executed by humans?*

ECONOMIC CONTEXT	Criteria	Description
SECTOR	Industrial sector	Which industrial sector is the system deployed in (e.g. finance, agriculture)?
BUSINESS FUNCTION & MODEL	Business function	What business function(s) is the system employed in (e.g. sales, customer service)?
	Business model	Is the system a for-profit use, non-profit use or public service system?
CRITICALITY	Impacts critical functions / activities	Would a disruption of the system's function / activity affect essential services?
SCALE & MATURITY	Breadth of deployment	Is the AI system deployment a pilot, narrow, broad or widespread?
	{Technical maturity}	*How technically mature is the system (Technology Readiness Level –TRL)*

DATA & INPUT	Criteria	Description
COLLECTION	Detection and collection	Are the data and input collected by humans, automated sensors or both?
	Provenance of data and input	Are the data and input from experts; provided, observed, synthetic or derived?
	Dynamic nature	Are the data dynamic, static, dynamic updated from time to time or real-time?
RIGHTS & IDENTIFIABILITY	Rights	Are the data proprietary, public or personal data (related to identifiable individual)?
	"Identifiability" of personal data	If personal data, are they anonymised; pseudonymised?
STRUCTURE & FORMAT	*{Structure of data and input}*	*Are the data structured, semi-structured, complex structured or unstructured?*
	{Format of data and metadata}	*Is the format of the data and metadata standardised or non-standardised?*
SCALE	*{Scale}*	*What is the dataset's scale?*
QUALITY AND APPROPRIATENESS	*{Data quality and appropriateness}*	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*

AI MODEL	Criteria	Description
MODEL CHARACTERISTICS	Model information availability	Is any information available about the system's model?
	AI model type	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?
	{Rights associated with model}	*Is the model open-source or proprietary, self or third-party managed?*
	{Discriminative or generative}	*Is the model generative, discriminative or both?*
	{Single or multiple model(s)}	*Is the system composed of one model or several interlinked models?*
MODEL-BUILDING	Model-building from machine or human knowledge	Does the system learn based on human-written rules, from data, through supervised learning, through reinforcement learning?
	Model evolution in the field $^{\text{ML}}$	Does the model evolve and / or acquire abilities from interacting with data in the field?
	Central or federated learning $^{\text{ML}}$	Is the model trained centrally or in a number of local servers or "edge" devices?
MODEL INFERENCE	*{Model development / maintenance}*	*Is the model universal, customisable or tailored to the AI actor's data?*
	{Deterministic and probabilistic}	*Is the model used in a deterministic or probabilistic manner?*
	Transparency and explainability	If information available to users to allow them to understand model outputs?

TASK & OUTPUT	Criteria	Description
TASKS	Task(s) of the system	What tasks does the system perform (e.g. recognition, event detection, forecasting)?
	{Combining tasks and actions into composite systems}	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*
ACTION	Action autonomy	How autonomous are the system's actions and what role do humans play?
APPLICATION AREA	Core application area(s)	Does the system belong to a core application area such as human language technologies, computer vision, automation and / or optimisation or robotics?
EVALUATION	*{Evaluation methods}*	*Are standards or methods available for evaluating system output?*

Note: Criteria and descriptions in grey and marked with an {} symbol = those where objective and consistent information is available. ML = for machine learning AI models.

More information: www.oecd.ai/classification | Contact: ai@oecd.org