

 **Deadline for public consultation comments: 30 June 2021**

OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS – PUBLIC CONSULTATION ON PRELIMINARY FINDINGS

Consultation at www.oecd.ai/p/classification

The [OECD Framework to Classify AI Systems](#) was developed by the OECD.AI Network of Experts to help policy-makers, regulators, legislators and others assess the opportunities and risks presented by different types of AI systems to inform their AI strategies. The Framework links the technical characteristics of AI systems with policy implications for the [OECD Principles on Trustworthy AI](#). After a year of work on the Framework, OECD.AI is launching a [public consultation](#) to gather feedback and input on the Framework’s usability and user-friendliness, and diverse perspectives and insights. All stakeholders are invited to partake in the consultation including: standards & technical bodies; business; legislators; regulator networks; civil society, consumers and others.

Key questions asked as part of the consultation:

1. Should **core and non-core criteria** be distinguished? I.e. should there be a **core classification framework for information that is generally accessible and additional, more complex or technical, considerations?**
2. Which **characteristics** should be **core criteria** and which ‘optional’?
3. Can AI systems be classified **consistently & reliably** with the core criteria?
4. Which criteria should be in a **more detailed, technically-oriented** framework?
5. Should there be **industry or application domain specific criteria** and classifications?

How can you participate?

1. By responding to an **online survey** to test the Framework on a real AI system (5 – 15 min).
2. By providing comments on the report or on one of the 4 key dimensions (10-30 min). Please send your comments to ai@oecd.org using the subject line: “Public consultation – comments on the OECD Framework for Classifying AI”.

Both the survey and report are available on: www.oecd.ai/p/classification

Before starting, please read the following carefully:

1. Unless you request otherwise, your name and input will be made available publicly: **please write “anonymous” in the subject line of your email** if you do not want your name and organisation to be published alongside your comment(s).
2. Please explain / provide the **reasons for your suggestions** to help us address them.
3. The present **report is not final** and will be subject to changes based on inputs.

Timeline for the consultation:

- **Deadline for sending comments / filling in survey:** 30 June 2021
- **Publication of comments:** July-August 2021

This project is part of the dedicated Programme on AI in Work, Innovation, Productivity and Skills (AI-WIPS), conducted by the OECD with the support of the German Federal Ministry of Labour and Social Affairs (BMAS).

Please send any queries about the consultation to: ai@oecd.org

We very much welcome your input – the Framework’s usefulness depends on it!

Table of Contents

OVERVIEW AND GOAL OF THE FRAMEWORK	4
Introducing the framework.....	4
Structuring elements.....	5
Applicability of the framework to a number of policy areas.....	7
I. CLASSIFICATION FRAMEWORK	9
1) CONTEXT	9
A. Industrial sector	9
B. Business function.....	10
C. Impacts critical functions / activities [optional criteria]	11
D. Scale of deployment and technology maturity	11
Breadth of deployment	12
AI system maturity [optional criteria]	12
E. Users of AI system.....	13
F. Impacted stakeholders, optionality and business model	13
Impacted stakeholders	13
Optionality	13
Business model: for-profit use, non-profit use or public service [optional criteria].....	13
G. Benefits and risks to human rights and democratic values.....	14
H. Benefits and risks to well-being [optional criteria]	15
Key AI actors in the “Context” dimension: operators of AI systems.....	15
2) DATA AND INPUT.....	16
A. Collection method, provenance and dynamic nature.....	16
Detection and collection of data and input	16
Provenance of data and input.....	16
Dynamic nature of data.....	17
Scale [optional criteria].....	18
B. Structure and format of data and input	18
Structure of data and input.....	18
Format of data and metadata [optional criteria].....	19
C. Rights and ‘identifiability’	19
Rights associated with data and input.....	19
‘Identifiability’ of personal data	20
D. Data quality and appropriateness [optional criteria].....	21
Key AI actors in the “Data and input” dimension: data collectors and data processors	22
3) AI MODEL.....	23
A. AI model characteristics	25
AI model type [core criteria]	25
Discriminative and generative models [optional criteria].....	26
B. Model building	26
Model building from machine-learned or human-encoded knowledge	26
Data interaction in machine learning models that evolve in the field.....	27
Central or federated learning [optional criteria]	27
C. Model inference, i.e. using a model [optional criteria].....	28

Key AI actors in the “AI model” dimension: developers and modellers	28
4) TASK AND OUTPUT	29
A. Task of the system [core criteria]	29
B. Action autonomy level [core criteria].....	30
C. Displacement potential	31
D. Combining tasks and actions into composite systems [optional criteria].....	31
E. Core application areas [core criteria].....	32
Key AI actors in the “Task and output” dimension: system integrator	32
II. APPLYING THE FRAMEWORK.....	33
Example 1: A credit-scoring system.....	33
A. Context	33
B. Data and input.....	34
C. AI model.....	35
D. Task and output	35
Example 2: AlphaGo Zero	35
A. Context	35
B. Data and input.....	36
C. AI model.....	37
D. Task and output	37
Example 3: Qlector.com LEAP system to manage a manufacturing plant	37
A. Context	38
B. Data and input.....	39
C. AI model.....	40
D. Task and output	40
Example 4: GPT-3	40
A. Context	40
B. Data and input.....	41
C. AI model.....	42
D. Task and output	42
III. ILLUSTRATIVE ETHICAL AND SOCIETAL RISK ASSESSMENT BASED ON THE FRAMEWORK	43
Overview and approach	43
Demonstration of risk mapping.....	43
Next steps	46
Annex A. Sample AI applications by sector, ordered by proxy for diffusion	47
Annex B. AI adoption per industry.....	49
Annex C. Expert meeting on classifying AI systems, Paris, 27 February 2020	51
Annex D. WG CAI Membership.....	52
Annex E. Meetings of the OECD Network of Experts Working Group on the Classification of AI Systems.....	55

OVERVIEW AND GOAL OF THE FRAMEWORK

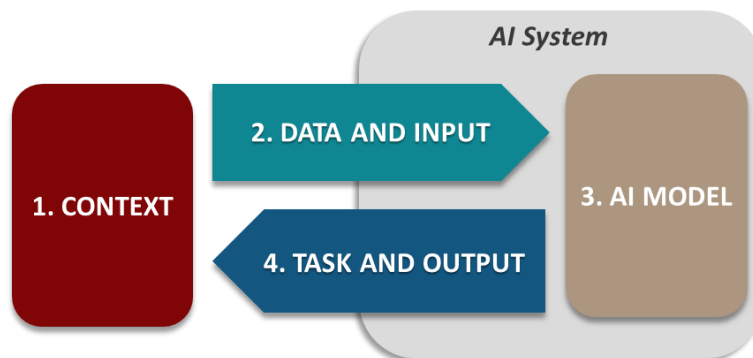
1. Different types of AI systems raise very different policy opportunities and challenges. The OECD AI Systems Classification Framework helps policymakers and others to classify AI systems according to their potential impact on public policy in areas covered by the OECD AI Principles. Section I provides a framework to assess AI systems' impact on policy. Section II puts the framework into use to classify particular AI systems. Section III illustrates how the framework can be used to assess basic risk associated with specific types of AI systems. This report helps to build a shared understanding of AI-related issues by mapping policy considerations to characteristics of specific types of AI systems.

2. The Framework is based on the work of the OECD Network of Experts' working group on the Classification of AI systems and the Secretariat. The expert group is co-chaired by Marko Grobelnik, AI Researcher & Digital Champion, AI Lab, Jozef Stefan Institute, Slovenia; Dewey Murdick, Director of Data Science, Center for Security and Emerging Technology (CSET), School of Foreign Service, Georgetown University, USA; and Jack Clark, co-chair of the AI Index at Stanford University, USA. Over sixty experts participate in the expert group (see Annex D). The working group held one physical and regular virtual meetings between February 2020 and March 2021 (Annex E).¹

Introducing the framework

3. The Framework classifies AI systems along four dimensions: 1) Context, 2) Data and input, 3) AI model and 4) Task and output (Figure 1). These dimensions build on the conceptual view of a generic AI system established in previous OECD work (Box 1).

Figure 1. Four dimensions of an AI System



1. The **context** describes the socio-economic environment in which the AI system is deployed and used. Core characteristics of this dimension include the sector in which the system is deployed (e.g., healthcare, finance, manufacturing) its business function; critical nature; scale of deployment; impacted stakeholders; user(s); impacts on human rights, and on well-being and its operator.

For example, systems used in healthcare raise specific patient data privacy considerations. In transportation, safety considerations are of paramount importance. AI systems used in public services will be held to high transparency and accountability standards, particularly in areas like security and law enforcement. AI systems in critical functions like energy

infrastructure raise specific security and robustness considerations. AI systems impacting groups like consumers raise consumer protection and product safety considerations.

2. **Data and input** refer to data and/or expert input (such as analytical functions and ontologies) that are used by the AI model to build a representation of the environment. Core characteristics of this dimension include provenance; collection method - by machines and/or humans; structure and format; and data properties (e.g., type, access).

Systems that use personal or sensitive data in the 'Data and input' dimension raise concerns about privacy, inclusiveness, human rights, and bias/fairness.²

3. An **AI model** is a computational representation of real world processes, objects, ideas, people and/or interactions that include assumptions about reality. Core characteristics of this dimension include the model characteristics; how the system is built (e.g., using expert knowledge, machine learning or both); and how it is used (e.g. for which objectives and using what performance measures).

Key properties of AI models, such as the degree of transparency and/or explainability, robustness, complexity, depend directly on the type of model, and the model building and inferencing processes. For example, systems using neural networks are often viewed as providing comparatively higher accuracy and less explainability than other types of AI systems. But explainability more generally is often tied to system complexity whereby the more complex a model is, the more difficult it is to explain.

The degree to which a model evolves in response to data is relevant to public policy and consumer protection regimes for AI systems that can iterate and evolve over time and may change their behaviour in unforeseen ways 'in the field'. In parallel, model evolution enables models to reflect the evolutions of the environments that they represent.

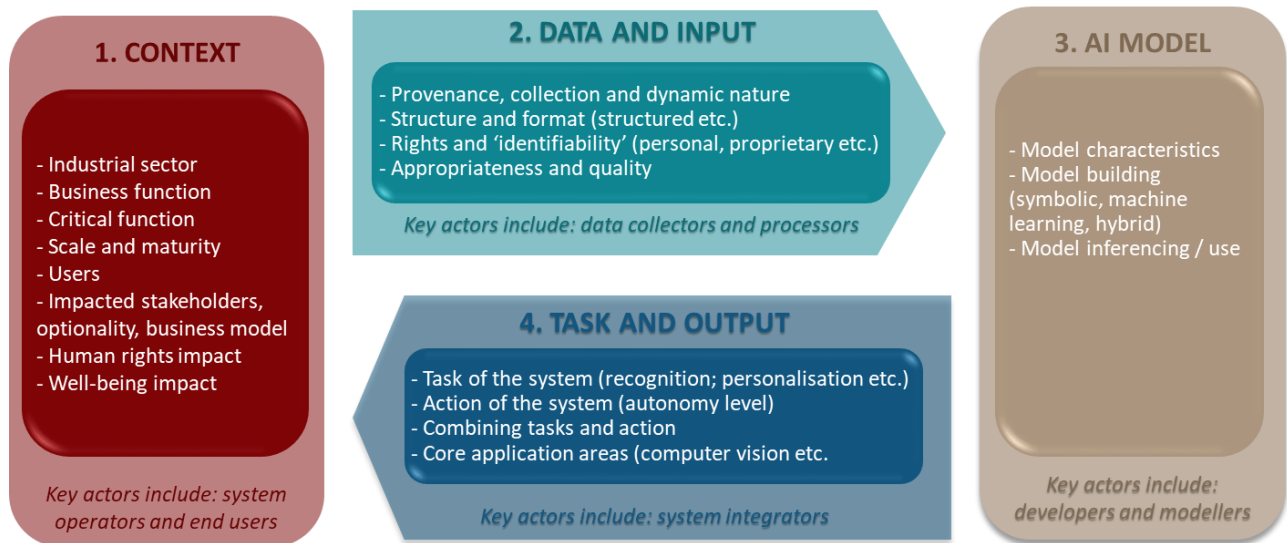
4. The **task and output** refer to the tasks the system performs e.g., personalisation or recognition and its outputs, as well as the resulting action that influences the context.³ Core characteristics of this dimension include system task(s); action autonomy; systems that combine tasks and actions like autonomous vehicles; and core application areas like computer vision.

For example, personalisation tasks generate outputs that could raise bias and fairness issues. Recognition tasks can raise concerns in relation to human rights, robustness and security, as well as bias. The tasks performed by AI systems also raise significant questions related to complementing or replacing human labour. The actions taken based on the outcomes of the AI system, e.g. by autonomous vehicles, also generate issues of fairness, safety, security and accountability.

Structuring elements

4. Each of the four dimensions has distinct properties and attributes, i.e. sub-dimensions that are relevant to assessing policy considerations associated with an AI system (Figure 2).

Figure 2. Characteristics per classification dimension and key actor(s) involved



Note: « key actors » are illustrative, non-exhaustive and based on the work of AIGO on the different stages of the AI system lifecycle, which are notably relevant to accountability

Source: based on the work of ONE AI and the AI system lifecycle work of AIGO (OECD, 2019⁽¹⁾)

5. The ten OECD AI Principles help structure the analysis of policy considerations associated with each dimension and sub-dimension. The Principles cover the following themes:

Values-based principles for all AI actors

Principle 1.1. People and planet

Principle 1.2. Human rights, privacy, fairness

Principle 1.3. Transparency, explainability

Principle 1.4. Robustness, security, safety

Principle 1.5. Accountability

Recommendations to policy makers for AI policies

Principle 2.1. Investment in R&D

Principle 2.2. Data, compute, technologies

Principle 2.3. Enabling policy and regulatory environment

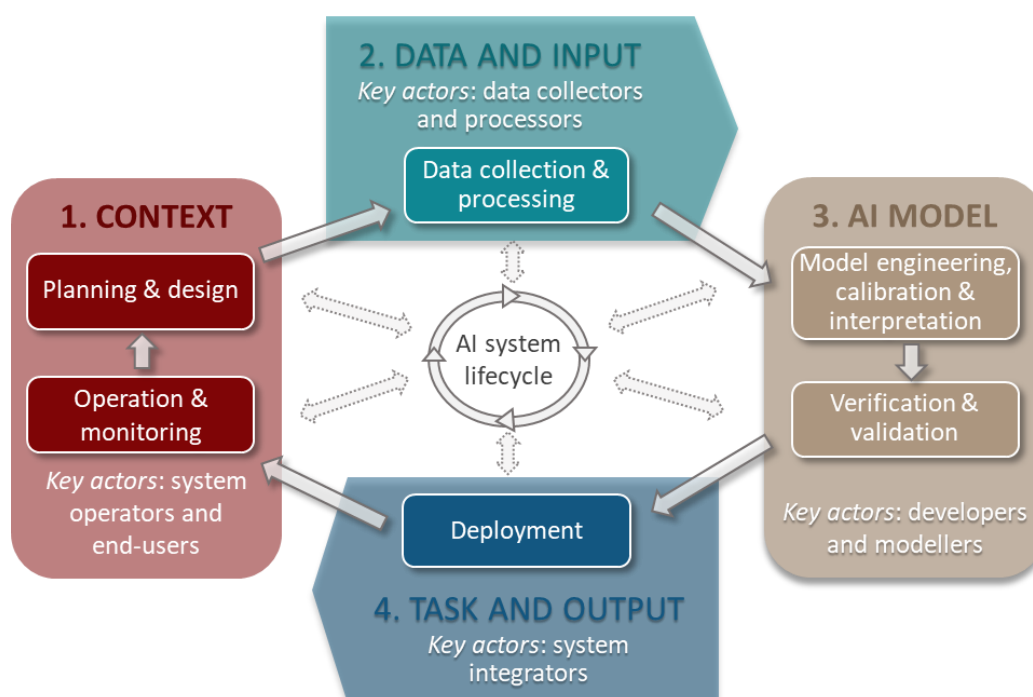
Principle 2.4. Jobs, automation, skills

Principle 2.5. International cooperation

6. The proposed classification aims to balance simplicity and user-friendliness with providing useful explanatory information. The expert group foresees the possibility of developing one “core framework” and one complementary, more detailed and technical framework containing “optional” characteristics.

7. The four dimensions of the classification framework can be associated with different stages of the AI system lifecycle (Figure 3) to identify relevant AI actors in the four dimensions, which are notably relevant to the accountability principle (Principle 1.5).⁴

Figure 3. The AI system lifecycle



Note: key actors are illustrative and not exhaustive and based on the work of AIGO on the different stages of the AI system lifecycle

Source: based on the AI system lifecycle work of AIGO (OECD, 2019^[11])

Applicability of the framework to a number of policy areas

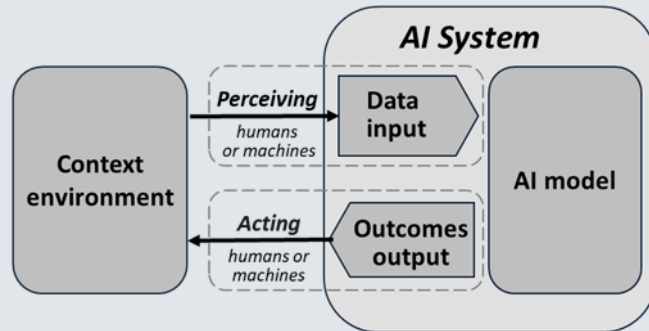
8. The framework's four dimensions can be used to consider a number of different policy issues, such as human rights and privacy (Principle 1.2); transparency and explainability (Principle 1.2); consumer protection policy, including consumer product safety (Principle 1.4); digital security and robustness (Principle 1.4); accountability by the actors involved in each phase (Principle 1.5)⁵; and international cooperation (Principle 2.5).

9. The area of jobs and skills for example, might include: 1) Context: job creation and displacement, access to training, dismissal policy, and social dialogue/worker consultation; 2) Data and input: job automation by sensors, new jobs in areas such as data science or data labelling; 3) AI model: building AI skills and attracting talent; and, 4) Tasks and output: task automation, job quality and quantity. These are being investigated in a complementary OECD project led by the Directorate for Employment, Labour and Social Affairs classifying the labour policy components and impacts of AI systems that is also part of the AI-WIPS programme.

Box 1. OECD characterisation of an AI System and its lifecycle, building on the OECD AI Principles (2019)

An AI system is a machine-based system that is capable of influencing the environment by producing recommendations, predictions or other outcomes for a given set of objectives.⁶ It uses machine and/or human-based inputs/data to: i) perceive environments; ii) abstract these perceptions into models; and iii) use the models to formulate options for outcomes.⁷ AI systems are designed to operate with varying levels of autonomy (OECD, 2019_[1]).

Figure 4. Stylised conceptual view of an AI system (per OECD AI Principles)



Source: (OECD, 2019_[1])

It should be noted that in the newer classification framework (Figure 1), the 'Perceiving' (data collection) and 'data / input' – that were separate objects in the original OECD work presented in this box – have been combined to simplify the framework (see dotted lines in Figure 4). The 'Outcomes' and 'Acting' objects have similarly been combined.

I. CLASSIFICATION FRAMEWORK

1) CONTEXT

10. The “Context” dimension of an AI system represents its broader socio-economic environment. This first dimension to classify an AI system is observable and can be influenced through actions resulting from an AI system’s outputs (OECD, 2019^[2]). Core characteristics of the “Context” dimension include the sector in which an AI system is deployed, its business function, its critical (or non-critical) nature, its deployment impact and scale, and its effects on human rights and well-being. The key AI actor in the “Context” dimension is the system operator.

Industrial sector

11. AI is diffusing rapidly and being applied in fields such as finance and insurance, advertising, transport, manufacturing and healthcare. Each industrial sector represents a different context that has implications in terms of industry structure, regulation, and policymaking for AI systems. In February 2020, the European Commission (EC) White Paper on AI (EC, 2020^[3]) called for a risk-based approach to AI regulation that would include risk linked to the sector, mentioning notably healthcare, transport, energy and parts of the public sector (*e.g.*, migration, border control, judiciary, social security, employment).

12. The Framework uses the International Standard Industrial Classification of All Economic Activities (ISIC REV 4). Using a standard industrial classification for the Framework allows for comparability with other sources of cross-country data on employment, skills, demography of enterprises, value-added and more. The highest-level sectoral categories of economic activities are called “sections”. They include:

- Section A Agriculture, forestry and fishing
- Section B Mining and quarrying
- Section C Manufacturing
- Section D Electricity, gas, steam and air conditioning supply
- Section E Water supply; sewerage, waste management and remediation activities
- Section F Construction
- Section G Wholesale and retail trade; repair of motor vehicles and motorcycles
- Section H Transportation and storage
- Section I Accommodation and food service activities
- Section J Information and communication
- Section K Financial and insurance activities
- Section L Real estate activities
- Section M Professional, scientific and technical activities
- Section N Administrative and support service activities
- Section O Public administration and defence; compulsory social security
- Section P Education
- Section Q Human health and social work activities
- Section R Arts, entertainment and recreation
- Section S Other service activities
- Section T Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use
- Section U Activities of extraterritorial organizations and bodies

13. The policy implications of deploying AI systems vary significantly from one sector to the next. Annex A provides an overview of key application areas in industrial sectors that are adopting AI most rapidly. Based on these key application areas, Table 1 provides a “heatmap” to illustrate the relevance of key policy considerations of AI for each industrial sector. The ranking is based on expert input.

AI systems are deployed in many different industrial sectors that raise different policy implications. The industrial sector has a particularly high impact on economic and social benefits (Principle 1.1) and on jobs and skills (Principle 2.4).

Table 1. Two-dimensional matrix flagging the relevance of the OECD AI Principles by industrial sector, based on survey of ONE AI experts


AI principle	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	2.5
Dimensions	Benefits people and planet	Human rights & fairness	Transparency & explainability	Robustness, security, safety	Accountability	Supporting Research	Data, compute, technologies	Enabling policy & regulatory environment	Jobs, automation, skills	International co-operation
1. CONTEXT										
1.a. Industrial sector										
Information and communication (Section J)		Very relevant	Very relevant	Very relevant	Very relevant	Low relevance	Very relevant	Low relevance	Relevant	Relevant
Professional, scientific and technical activities (Section M)	Relevant	Very relevant	Relevant	Low relevance	Relevant	Very relevant	Very relevant	Low relevance	Very relevant	Low relevance
Financial and insurance activities (Section K)	Relevant	Very relevant	Very relevant	Very relevant	Relevant	Low relevance	Low relevance	Very relevant	Relevant	Relevant
Administrative and support service activities (Section N)		Relevant	Very relevant	Low relevance	Relevant	Low relevance	Low relevance	Low relevance	Relevant	Low relevance
Agriculture, forestry and fishing (Section A)	Very relevant	Low relevance		Low relevance	Low relevance	Very relevant	Relevant	Low relevance	Low relevance	
Manufacturing (Section C)	Low relevance	Relevant		Very relevant	Low relevance	Low relevance	Low relevance	Very relevant	Very relevant	
Public administration and defence (Section O)	Very relevant	Very relevant	Very relevant	Very relevant	Very relevant	Very relevant	Very relevant	Low relevance	Relevant	
Wholesale and retail trade (Section G)		Relevant	Relevant	Relevant	Relevant		Low relevance	Very relevant	Very relevant	
Education (Section P)	Very relevant	Very relevant	Very relevant	Relevant	Very relevant	Very relevant	Relevant	Low relevance	Low relevance	
Human health and social work activities (Section Q)	Very relevant	Very relevant	Very relevant	Very relevant	Very relevant	Very relevant	Very relevant	Very relevant	Very relevant	Relevant
Transportation and storage (Section H)	Relevant	Low relevance	Relevant	Very relevant	Relevant	Low relevance	Low relevance	Very relevant	Very relevant	Relevant
Electricity, gas, steam and air conditioning supply	Very relevant	Low relevance		Very relevant	Low relevance	Low relevance	Low relevance	Very relevant	Low relevance	Low relevance
Water supply; sewerage, waste management and remediation activities	Very relevant	Low relevance		Very relevant	Low relevance	Low relevance	Low relevance		Low relevance	Low relevance


Note: A Principle is considered to be "very relevant" if it is deemed essential to obtain beneficial outcomes from AI in the sector considered; "relevant" if it should be considered when deploying AI in a sector, but not as the first-order priority; and has "low relevance" if it can potentially contribute to obtaining beneficial outcomes in the sector considered, but is not necessary. Cells are blank when an AI Principle is regarded as not applicable to the sector considered. A consultation process is being conducted to collect additional expert input through a survey on the relevance of each AI Principle by industrial sector.

Source: OECD

B. Business function


14. Functional areas in which AI systems can be employed include but are not limited to:

-  Human resource management;
- Sales;
- ICT management and information security;
- Marketing and advertisement;

- Logistics;
- Citizen/customer service;
- Procurement;
- Maintenance;
- Accounting;
- Monitoring and quality control;
- Production;
- Planning and budgeting;
- Research and development  and
- Compliance and justice.

Different AI systems can perform the same task in different functional areas, with different implications for policy making. For instance, a forecasting algorithm used to improve (optimise) logistics may have different implications than a forecasting system designed to support hiring decisions. The business function for which the AI system is used will have a particular impact on economic and social benefits (Principle 1.1); fairness and absence of bias (Principle 1.2); security, safety and robustness (Principle 1.4) and jobs and skills (Principle 2.4)

C. Impacts critical functions / activities [optional criteria]

15. Activities that are considered “critical” are economic and social activities of which “the interruption or disruption would have serious consequences on: 1) the health, safety, and security of citizens; 2) the effective functioning of services essential to the economy and society, and of the government; or 3) economic and social prosperity more broadly” (OECD, 2019_[2]); (OECD, 2019_[4]). It is important to note that not all systems in a critical sector are critical. For example, the administrative time tracking systems of a hospital or a bank are not considered to be critical systems. 

- *AI system is in a critical sector or infrastructure*: an AI system is deployed in a critical sector or infrastructure (e.g., energy, transport, water, health, digital infrastructure and finance).
- *AI system performs a critical function independent from its sector*: an AI system serves a critical function independent from the sector (e.g., conducting elections, maintaining supply chains, law enforcement, providing medical care, supporting the financial system).


In some sectors critical functions are accompanied by heightened risk considerations with ex ante regulations. The critical function will have particular impact on security, safety and robustness (Principle 1.4). In the European Union, the NIS Directive mandates the supervision of critical sectors. EU Member states must supervise the cybersecurity of critical market operators ex-ante in critical sectors like energy, transport, water, health, digital infrastructure and finance sector, and ex-post surveillance is required for critical digital service providers like online market places, cloud services and online search engines. In the United States, national critical functions include conducting elections, maintaining supply chains, law enforcement, and providing medical care (CISA, 2019_[5]). In the financial sector, banks operate some critical functions, like the SWIFT system for sending payment orders between financial institutions.

D. Scale of deployment and technology maturity

16. AI systems’ economic and social impact vary depending on the following four factors: the breadth of deployment of an AI system; its maturity; the stakeholder(s) impacted by the system, and; the for-profit use, non-profit use or public service use of the system.


Breadth of deployment

17. The breadth of deployment relates for example to the number of individuals that are or will be affected by a system.

- *A pilot project:*
- *Narrow deployment:* the deployment could for example be at the level of one company or of one country
- *Broad deployment:* the deployment could for example be at the level of one sector
- *Widespread deployment:* the deployment could for example reach across countries and sectors 

AI system maturity [optional criteria]


18. The maturity of the deployed AI system can vary widely. Technology readiness levels (TRLs) can help classify AI technologies' maturity. The following categories are based on JRC analysis (Martinez Plumed, 2020^[6]) building on the NASA TRL framework (Mankins, 1995^[7]).

- *Basic principles observed and reported – TRL 1:* this is the lowest level of technology readiness where research begins to be translated into applied R&D. Sample output might be a scientific article on a new technology's principles.
- *Technology concept and/or application formulated – TRL 2:* speculative practical applications are invented based on assumptions not yet proven or analysed. Sample output might be a publication or reference highlighting the applications of the new technology 
- *Analytical and experimental critical function and/or characteristic proof-of-concept – TRL 3:* Continued research and development efforts include analytical studies and lab studies to physically validate analytical predictions of separate elements of the technology. Sample output might be measurement of parameters in the lab.
- *Component and/or layout in controlled environment – TRL 4:* Basic technological components are integrated to verify they can work together but in a relatively superficial manner. Sample output might be integration of "ad hoc" software or hardware in the laboratory.
- *Component and/or layout validated in relevant environment – TRL 5:* Reliability is significantly increased and basic technological components integrated with fairly realistic supporting elements that can be tested in a simulated environment. Sample output might be realistic laboratory integration of components.
- *Representative model or prototype system demonstrated in relevant environment – TRL 6:* Sample output might be testing a prototype in a realistic laboratory environment or in a simulated operational environment.
- *System prototype demonstration in operational environment – TRL 7:* Examples include testing the prototype in operational testing platforms (e.g., a real-world clinical setting, a vehicle, etc.).
- *System or subsystem complete and qualified through test and demonstration – TRL 8:* Technology proved to work in its final form and under expected conditions. In most cases, this TRL represents the end of true system development. Examples include developmental test and evaluation of the system to determine if the requirements and specifications are fulfilled.
- *Actual system or subsystem in final form in operational environment – TRL 9:* Actual application of the technology conditions such as those encountered in operational conditions. Strong monitoring and improvement processes are critical to continue to improve the system.

AI system maturity is particularly relevant to safety, robustness and security (Principle 1.4), accountability (Principle 1.5) and R&D investment (Principle 2.1).

E. Users of AI system

19. The users of an AI system are often not those who developed, implemented or operate it. Users can range in competency from AI expert(s) to amateur end-user(s). For example, AI systems deployed in sectors such as healthcare or agriculture are often used by practitioners or domain experts who are not typically AI experts. In light of the implications of the level of end-user expertise, AI systems can be distinguished based on whether their typical users have any systems operation training:

- *Amateur*: a user who has no training
- *Trained practitioner who is not an AI expert*: a user with some specific training on how to use the AI system in question 
- *AI expert practitioner*: a user with specific training and knowledge of AI (an AI expert or system developer)


The users of an AI system relate notably to accountability (Principle 1.5); transparency and explainability (Principle 1.3); and safety, security, and robustness (Principle 1.4).

F. Impacted stakeholders, optionality and business model

Impacted stakeholders


20. Stakeholders impacted by the system may be:

- Consumers
- Workers / employees
- Business
- Government agencies / regulators
- Specific communities
- Children or other vulnerable or marginalised groups

The stakeholders impacted by the system are most relevant to transparency and explainability (Principle 1.3) and to policy and regulatory frameworks (Principle 2.2). Stakeholder groups such as consumers, workers/employees, or children are often covered by existing policy and regulatory regimes. In Europe the GDPR gives data subjects the right to not be subject to automated decision-making 

Optionality



21. Optionality or dependence refers to the degree of choice that users have on whether to be subject to the effects of an AI system or not. Optionality can be understood as the extent to which users can opt out of 'the effects' or 'the influence' of the AI system, e.g. by switching to another AI system, also referred to as “switchability” (AI Ethics Impact Group, 2020^[8]) and the societal repercussions of doing so, e.g. for access to healthcare or financial services.

- Users cannot *opt out* of the AI system’s output.
- Users can *opt out* of the AI system’s output.
- Users can  *override or correct* the AI system’s output.

Business model: for-profit use, non-profit use or public service [optional criteria]

22. Operators may be using an AI system (OECD, 2011^[9]):

- For-profit use – subscription fee model



- For-profit use – advertising model
- For-profit use – other model
-  Non-profit use (outside public sector) – voluntary donations and community models
-  Public service
- Other

The type of use and business model of the AI system operator can be relevant to determining the objectives that an AI system is optimising for, e.g. maximising engagement in advertising-based business models.

G. Benefits and risks to human rights and democratic values

23. Some AI systems generate outputs that can impact individuals’ human rights (see Table 2). Low-risk contexts for individuals, such as a restaurant recommendation, may not need multi-layered and costly approaches, and can therefore rely mostly on machines, if there are adequate protections regarding combining different data sets at aggregate levels.

Table 2. Potential impact on select human rights and democratic values

AI outcomes impact human rights or democratic values:	Outcome-dependent	No impact
Human dignity, life and physical and mental integrity		
Liberty and security		
Fair trial; no punishment without law; effective remedy		
Privacy and family life, and the protection of personal data 		
Freedom of thought, conscience and religion		
Freedom of expression; assembly and association		
Non-discrimination and equal treatment		
Protection of property and peaceful enjoyment of possessions		
Right to education 		
Right to democracy and free elections		
Rights of the child, the elderly and persons with disabilities		
Other (detail)		

Note: International human rights refer to a body of international laws, including the Universal Declaration of Human Rights,¹ as well as regional human rights systems such as that of the Council of Europe. Human rights provide a set of universal minimum standards based on, among others, values of human dignity, autonomy and equality, in line with the rule of law. Human rights overlap with wider ethical concerns and with other areas of regulation relevant to AI, such as personal data protection or product safety law. Risk assessment frameworks would usually also include the likelihood the risk will occur, its impact and mitigation measures.

Source: based on (CoE, 2020_[10]), (CoE, 1998_[11]) and (OHCHR, 2011_[12]).

Transparency and explainability (Principle 1.3) as well as accountability (Principle 1.5) are widely viewed as having higher importance in contexts where the outcomes of an AI system can impact human rights. Examples include AI used to sentence criminals, recommend decisions about educational opportunities or conducting job screening. Such high-stakes situations often require formal transparency and accountability mechanisms (Principles 1.3 and 1.5), including transparency about the role of AI, human involvement in the process (e.g., “human in-the-loop”) and the availability of appeals processes, particularly where life and liberty are at stake. Across the spectrum, people broadly agree that AI-based outcomes (e.g., a score) should not be the only decisive factor when applications or decisions have a significant impact on people’s lives. Such applications may for example require that a human consider the social context so as to help avoid unintended consequences. For example, the GDPR stipulates that a human must be in the loop if a decision has legal or similarly significant effects on people.⁸

H. Benefits and risks to well-being [optional criteria]

24. Many AI systems generate outputs that can impact individuals' well-being either positively or negatively. This impact pertains to different areas of life, such as work and job quality, environment quality, social connections and civic engagement, among others (Table 3).

Table 3. AI outcomes' impact on well-being

AI outcomes impact well-being:	Outcome-dependent	No impact
Health (including mental health)		
Housing		
Income and wealth		
Work and job quality		
Environment quality		
Social connections		
Civic engagement		
Education		
Subjective well-being		
Work-life balance		

Note: captured for example by measures of inequality

Source: (OECD, 2020^[13])

Key AI actors' in the "Context" dimension: operators of AI systems

25. The Context dimension can be associated with the 'planning and design' stage of the AI system lifecycle as well as, following deployment, with the 'operation and monitoring' phase. Planning and design of the AI system involves articulating the system's concept and objectives, underlying assumptions, context and requirements (OECD, 2019^[14]). Planning and design currently involves expertise such as data scientists, domain experts, and governance experts.

26. 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 the system's 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. AI systems operators can establish:

- *Transparent information about the AI system's objectives and assumptions:* providing interested stakeholders with access to useful information.
- *Performance monitoring mechanisms:* i.e., metrics to assess the performance and accuracy of the AI system.
- *Tools or processes for trustworthy AI:* using tools like guidelines; governance frameworks; product development or lifecycle tools; risk management frameworks; sector-specific codes of conduct; process standards; technical validation approaches; technical documentation; technical standards, toolkits, toolboxes or software tools; educational material; change management processes; or certification (technical and/or process-related).

The 'context' dimension relates to accountability (Principle 1.5); transparency and explainability (Principle 1.3); and safety, security, and robustness (Principle 1.4). Key actors are often AI system operators.

2) DATA AND INPUT

27. An AI system can be based on expert input and/or on data, which can be generated by humans and/or automated tools (e.g., machine learning algorithms). Historically, AI systems were powered by *expert input* in the form of logical representations that formed the basis of early optimisation and planning tools such as those used in medical diagnosis, credit card fraud detection or in chess playing (e.g., IBM’s Deep Blue”). They required that researchers build detailed decision structures to translate real-world complexity into rules to help machines arrive at human-like decisions. Expert input also includes structures such as ontologies, knowledge graphs, decision rules, and analytical functions (see section on Structure of data).¹⁰ Over recent years, AI systems have become increasingly statistical and probabilistic and are powered by a growing variety of data types.

28. Core characteristics of the ‘data and input’ dimension relate to provenance, the data collection and origin (e.g., data collection, origin, dynamic nature, and scale); their technical characteristics (e.g., structure and encoding); domain (e.g., personal, proprietary or public); and data quality and appropriateness. The next sections draw on the OECD 2019 work on “Enhancing Access to and Sharing of Data.”

A. Collection method, provenance and dynamic nature


Detection and collection of data and input


29. Data and input can be detected and collected (“tracked”) from the environment by humans or machines (also known as “telemetry”):

- *Collected by humans*: a human may for example be needed to observe and collect information on someone’s mental state. Other examples of data collected by humans are crowd sourcing data and human-based computation, where a machine outsources certain steps of the computation process to humans.
- *Collected by automated sensors*: devices that automatically monitor and record data include cameras, microphones, thermometers, laboratory instruments and other sensors such as Internet of Things (IoT) devices, but also automated recording of information from online log files, mobile phones, GPS watches and activity wristbands.
- *Collected by humans and automated sensors*: some data are collected by humans together with automated tools e.g., in healthcare applications, data from sensors such as heartbeat or blood pressure detectors will often be combined with a doctor’s assessment.

Data and input collection through sensing can benefit society in fields such as health care and safety (e.g., activity trackers associated with health applications) or environmental applications. Data and input collection can also bring labour market considerations (Principle 2.4), including the automation of tasks (e.g., security surveillance or maintenance assessments); improving worker safety; measuring worker productivity; and codifying expert knowledge.

Provenance of data and input

30. Several categories of data provenance (or “origin”) can be distinguished. The following draws on the categorisation made by Abrams (2014)¹¹ and (OECD, 2019_[15]) on data being collected from individuals with decreasing levels of awareness. It should be noted that these categories can overlap. In the present section, the original categorisation is broadened to also cover expert input and non-personal data, as well as data that are synthetically generated 

- *Expert input*: expert input is typically human knowledge that is codified into rules and structures such as ontologies, knowledge graphs, and analytical functions (e.g., the objective function or rewards an AI model will optimise for).
- *Provided data*: data that originate from actions by individuals or by organisations that are aware of the data being provided. They include initiated (e.g., a license application), transactional (e.g., bills paid), and posted (e.g., social networking posts) data.
- *Observed data*: observed data are collected through observation of a behaviour or activity through human observation or the use of automated instruments or sensors. Examples include website visitor provenance and browsing patterns observed by a website administrator. Observed data also include data such as sounds, scents, temperature, GPS position or soil acidity. Observed data about individuals can be engaged (e.g., accepting cookie tracking on a website), unanticipated (e.g., the tracking of seconds spent looking at a specific image online) or passive (e.g., CCTV images of individuals).
- *Synthetic data*: synthetic data are usually generated by computer simulations. Synthetic data allow for simulation of scenarios that are difficult to observe or replicate in real life (e.g., a car accident) or are otherwise too expensive to collect at scale (e.g., millions of miles of driving time for self-driving cars). They include most applications of physical modelling, such as music synthesisers or flight simulators. The output of such systems approximates reality, but is generated algorithmically.
- *Derived data*: data that are derived from other data to become a new data element. Derived data include computational (e.g., a credit score) and categorical data (e.g., age group of a buyer). Derived data can be inferred (e.g., the product of a probability-based analytic process like a fraud score or risk of accident) or aggregated (e.g., abstracted from more fine-grained data 

Awareness and consent for the provision of personal data about individuals is a critical focus area for privacy and consumer policy.

Synthetic data allow for simulation of scenarios that are difficult to observe or replicate in real life (e.g., a car accident). Expert input is typically human knowledge that is codified into rules.

Dynamic nature of data

31. Data can be static or dynamic to varying extents:

- *Static data*: static data do not change after they are collected (e.g., a given publication, a product's batch number, or the geographic latitudes and longitudes of a fixed element like a building or a mountain).
- *Dynamic data updated from time-to-time*: dynamic data continually change after they are recorded in order to maintain their integrity. Models relying on dynamic data can leverage “incremental algorithms,” that update the model frequently based on incoming data. Dynamic data can be updated from time-to-time without necessarily being real-time. Examples include timetables of flights' estimated time of arrival using batch processing.
- *Dynamic real-time data*: Dynamic real-time data are delivered immediately after collection with no delay. Examples of systems that use real-time data processing include an alarm system triggered by an entry signal; a recommender system that evolves in real-time as it is being used (e.g., with a streaming video service like YouTube); and an autonomous driver system that reacts to real-time environmental data.

The degree to which data are static or dynamic is particularly relevant to public policy for AI systems that can iterate and evolve over time and may change their behaviour in unforeseen ways.

Scale [optional criteria]

32. The scale of a data set is a continuous variable that has an ever increasing upper bound. If real-time, scale can be roughly measured as the order of magnitude of bytes per time unit (e.g., tens of petabytes per second) or the number of requests (to the AI System) per second; if static, size as measured in bytes, (e.g., hundreds of gigabytes). The scale of data continues to change as technology advances. The upper bound is generally reached by very few government and commercial enterprises that accommodate high-velocity, real-time data streams at extremely large data volumes.

33. The scale of data can be:

- *Very large*: one exabyte or larger. Extremely large volumes of data take time to gather/accumulate and require complex systems to operate and process and process.
- *Large*: tens of petabytes per second (if real-time).
- *Medium*: hundreds of gigabytes (if static).
- *Small*: tens of gigabytes or smaller. There are no constraints to transfer and process ‘small’ data in current broadband networks and computing environments.

AI powered by machine learning techniques is known to rely on large volumes of data to function well, based on which patterns are inferred. There is active research on AI systems that use less data, such as one-shot learning (Principle 2.1). These AI systems are learning through self-play – via reinforcement learning – to drastically reduce the scale of the data needed to train a model.


In terms of robustness of AI systems (Principle 1.4), researchers have found that there is a trade-off between the quantity of data and the number of variables in a model. A larger model – one with more parameters – needs less data to achieve the same performance as a smaller one. This has implications for problems where training data samples are expensive to generate and likely confers an advantage to large companies entering new domains with models based on supervised learning.¹²

Data size also relates to building the technology infrastructure to process, transfer and share large volumes of data for AI (Principle 2.2).

B. Structure and format of data and input

Structure of data and input

34. This sub-dimension identifies the different types of data structures:

- *Unstructured data*: include data that either do not have a pre-defined data model or are not organised and labelled in a pre-defined manner (e.g., text, image, audio, video and interlinkages between graph networks, social media, or website data). Unstructured data often include irregularities and ambiguities that are difficult for traditional programs to analyse.  Unstructured data are sometimes referred to as “raw data”.
- *Semi-structured data*: in practice, most data combine both unstructured and structured data. For example, a photo taken with a smart phone consists of the image itself (unstructured data), accompanied by structured metadata about the image (when and where it was taken, what device took the picture, the picture format, its resolution, etc.). Similarly, data on a social network such as Twitter includes unstructured text, alongside structured metadata about the author of the text and his or her networks. Examples of semi-structured data include social media and device or sensor data.
- *Structured data*: structured data are stored in a pre-defined format and are straightforward to analyse. Structured data have labels describing their attributes and relationships with other data. Vast amounts of user data from websites or e-commerce sites are structured, fuelling the

development of a wide variety of marketing techniques (e.g., personalised advertisements), recommendation mechanisms (e.g., Amazon products, Netflix content, Spotify recommendations, YouTube 'up next') and engagement systems (e.g., Facebook feeds). Examples of structured data include interlinked tables and databases.

- *Complex structured data*: complex structured data are often produced in the form of a model, which is both the output of an AI algorithm and can be used as input to another system. Examples of complex structured data include ontologies (e.g., partial models of the environment); knowledge graphs; rules (e.g., expert systems); and analytical functions (e.g., adversarial learning or reinforcement learning functions).

35. Each of these structural alternatives can be encoded in various data types (e.g., binary, numeric, text) and represent different types of media (e.g., audio, image, video, or combinations of media types). Data labeling is the process of tagging data samples, which generally required human knowledge to build training data.

Data and input structure relates to transparency and auditability (Principle 1.3) and directly impacts the AI model choice. Structured data is easier to document and audit (Principle 1.5) and also influences data sharing policies (Principle 2.2)

Format of data and metadata [optional criteria]

36. *Data format* (or “encoding”) refers to the format of the data themselves. Data format is closely related to data collection (e.g., a camera will produce a specific format of image data) and to data modelling, where different modelling techniques will require specific data formats (e.g., time series modelling requires temporally sequenced data).

37. *The dataset metadata* may include information on how a dataset was created, its composition, its intended uses and how it has been maintained over time. Formats / standards for annotating datasets often need to be developed by sector and by type of use.

- *Standardised data format*: standardised data have a pre-agreed format, which enables a data set to be compared to other data sets.
- *Non-standardised data format*: data can also be in “ad hoc” formats created for the purpose of particular applications (e.g., video).
- *Standardised dataset metadata*.
- *Non-standardised dataset metadata*.

Standardisation of data formats facilitates interoperability and data re-use across applications, accessibility and can help ensure that data are findable, catalogued, searchable and re-usable. The use of standardised formats may improve a system’s robustness and security by making it easier to address security vulnerabilities (Principle 1.4).

In some sectors, standardised templates for dataset metadata annotation are being developed. Standardised metadata facilitates the development and sharing of training datasets and by extension can help accelerate the development and use of AI systems (Principle 2.2).

C. Rights and ‘identifiability’

Rights associated with data and input

38. This sub-dimension distinguishes the rights (also called “domain”) associated with data and input used by an AI system, which entail different policy implications when used in training data and/or in

deployment context. Data domains include the following three categories, which can overlap in certain applications (see Figure 5):

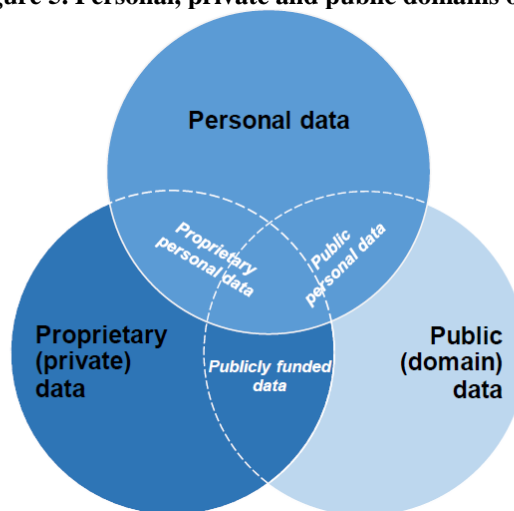
- *Proprietary data*: proprietary data are data that are privately held, often by corporations, and are typically protected by intellectual property rights – including copyright and trade secrets – or by other access and control regimes, such as contract and cyber-criminal law. There is typically an economic interest to restrict access to proprietary data.¹³ Input into an AI system in the form of rules can be considered as a form of proprietary data.
- *Public data*: public data are not protected by intellectual property rights or any other rights with similar effects and in many cases can be shared for access and re-used through open data regimes.
- *Personal data*: personal data is data that “relates to an identified or identifiable individual”.

Proprietary data raise issues such as transparency and explainability (Principle 1.3) and bias in AI systems (Principle 1.2), as well as considerations of business scale-up (Principle 2.2).

Public data is relevant to economic, social and environmental impacts (Principle 1.1); research (Principle 2.1); and data availability and compute capacity (Principle 2.2).

Personal data is associated with privacy considerations and legislation, and usually requires more restrictive access regimes. Personal data is relevant to issues related to human rights, fairness and privacy (Principle 1.2).

Figure 5. Personal, private and public domains of data



Source: (OECD, 2019^[16]).

'Identifiability' of personal data

39. Personal data taxonomies have been introduced to differentiate between different categories of personal data. ISO/IEC 19441 (2017) distinguishes five categories or “states” of data identifiability:

- *Identified data*: identified data are data that can be unambiguously associated with a specific person because they contain personal identifiable information.
- *Pseudonymised data*: pseudonymised data are data for which all identifiers are substituted by aliases. The alias assignment is such that it cannot be reversed by reasonable efforts, except for the party that performed the assignment.

- *Unlinked pseudonymised data*: unlinked pseudonymised data are data for which all identifiers are irreversibly erased or substituted by aliases. The linkage cannot be re-established by reasonable efforts, including by the party that performed the assignment.
- *Anonymised data*: anonymised data are data that are unlinked and of which the attributes are altered (e.g., the attributes' values are randomised or generalised) in such a way that there is a reasonable level of confidence that a person cannot be identified, directly or indirectly, by the data alone or in combination with other data.
- *Aggregated data*: aggregated data are statistical data that do not contain individual-level entries and are combined with information about enough different persons that individual-level attributes are not identifiable.

The type of personal data used by AI systems has implications for individuals' human rights, fairness and privacy (Principle 1.2). Data identifiability can help assess the level of risk to privacy and informs the need for legal and technical protection and access control. Concerns are raised that even absent personal data, AI systems are able to infer data and correlations from proxy variables that are not personally identified, such as purchasing history or location.

In addition, some regimes such as the EU's GDPR distinguish 'sensitive personal data' that consist of racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, health data or data concerning a person's sex life or sexual orientation. In the US, personal data considered sensitive include data about children, financial and health information.¹⁴

D. Data quality and appropriateness [optional criteria]

40. Data appropriateness (or “qualification”) is about defining criteria to ensure that the data are appropriate for use in a project *i.e.* fit for purpose and relevant to the system or process. For example, in clinical trials to evaluate drug efficiency, criteria for using patient data must include patients' clinical history (previous treatments, surgery, etc.). Data quality also plays a key role for AI systems.

- *Data appropriateness*: criteria exist and ensure that data are appropriate for the purpose for which they are to be used.
- *Sample representativeness*: the selected variables and the training or evaluation data are able to accurately depict / reflect the population in the AI system environment.
- *Adequate sample size*: the sample size displays an appropriate level of granularity, coverage, and sufficiency of data.
- *Completeness and coherence of sample*: the sample is complete with low missing or partial values. Outliers must not affect the quality of data either.
- *Low data “noise”*: *i.e.* the data is infrequently incorrect, corrupted or distorted (e.g., intentional or unintentional mistakes in survey data; data from defective sensors).

Data appropriateness impacts the accuracy and reliability of the outcome of AI systems and relates to their robustness, security and safety (Principle 1.4.) The use of inappropriate data/input in an AI system can lead to erroneous and possibly dangerous conclusions.¹⁵

Data quality has important policy implications to human rights and fairness (Principle 1.2), as well as to the robustness and safety of AI systems (Principle 1.4): from both fairness and robustness perspectives, datasets must be inclusive, diverse and representative so they do not misrepresent specific (sub) groups.

Key AI actors in the “Data and input” dimension: data collectors and data processors

41. The Data and input dimension maps directly to the ‘data collection and processing’ stage of the AI system lifecycle (Figure 3), which includes gathering and cleaning data, possibly labelling, 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 (OECD, 2019_[14]). Data collection and processing currently involves expertise such as data scientists, domain experts, data engineers, and data providers. Actions performed by data collectors and processors include:

- *Performing checks:* for data quality and appropriateness.
- *Transparent information about the data and inputs used in the AI system:* providing interested stakeholders with access to useful information on the data and inputs used in the AI system.
- *Labelling data:* i.e., tagging data with informative data.
- *Protecting personal data.*
- *Documenting data and dataset characteristics.*
- *Using tools or processes for trustworthy AI:* data collectors and processors may use tools like guidelines, governance frameworks, Product development / lifecycle tools, Risk management, Sector-specific codes of conduct, Process standards, Technical validation approaches, Technical documentation, Technical standards, Toolkits / toolboxes / software tools, Educational material, Change management processes, Certification (technical and/or process-related).

3) AI MODEL

42. This dimension considers AI models as composites of multiple core technical components, and analyses the choice of AI models, how the models are built, how the models are interlinked with one another and other sub-systems, and how they are used (model inferencing). This section also discusses how these AI model traits relate to policy considerations.

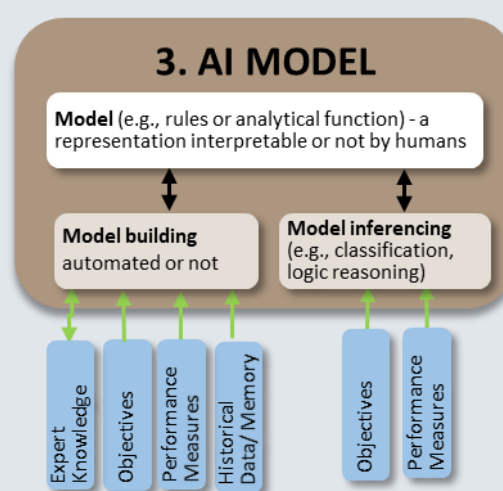
43. Models can be based on expert knowledge and/or data and can be processed by humans and/or by automated tools (e.g., machine learning algorithms) (Box 2).

Box 2. OECD characterisation of an AI Model included in the OECD AI Principles (2019)


To cover different types of AI systems and different scenarios, Figure 6 separates:

- The *AI model* itself, an object that forms the core of an AI system and represents all or part of the system's external environment.
- The *model building* process, often called "training" or "optimisation".
- The process of *using the model*, in which model inferencing algorithms generate outputs for information or action from the model given specific objectives and performance measures.

Figure 6. Detailed conceptual view of an AI Model



Source: (OECD, 2019₍₁₁₎)

44. **What is an AI model?** AI models are computational representations of processes, objects, ideas, people, and/or interactions taking place in the context (or the environment). AI models follow human-designed procedures to represent, and interact with, real or artificial world processes 

45. **What is the purpose of an AI model within a deployed system?** To accurately classify AI systems, it is helpful to identify the core AI model, or models, around which the system is built. The core models used in an AI system determine a wide array of characteristics. When classifying an AI model, it can be useful to analyse the *purpose* of the model, i.e. the problem that the AI system is trying to solve. The model choice and model building depend on the purpose of the AI system. AI models can be used to make recommendations in response to an input, e.g. answering a question ("what is the best next move in a chess game?"), generate data in response to a prompt ("what would a person look like if they were 20 years older?"), make predictions about the future courses of action ("will this road become congested?") and a wide range of other processes (see A. Task of the system [core criteria]).

46. **AI models are not universal:** It is important to highlight that AI models include assumptions about the world around them and several possible types of bias (**Error! Reference source not found.**). Many different representations of the same system can be developed to serve different purposes, i.e., there is no 'universal' – unique or "correct" – model to represent a given reality.

47. **One model, or many models? Most AI applications use composite models e:** In practice, the vast majority of AI systems in deployed, real-world contexts are *composite systems*. They are composed of a variety of interlinked AI sub-models derived from different sources, working together for a specific purpose. For instance, Google Photos is a popular application for storing and searching photos. Google photos consists of several (sub)systems, including a system for storing data (photo storage), a system for searching data via words that the user specifies, and a system for interpreting the search terms which leverages additional systems, some of which are AI models. When a user searches for photos by date, a human-designed system looks up photos' upload date, or the date encoded in the photo metadata. When a user searches for photos by a concept such as "flowers", an AI model carrying out a recognition function tries to match photos with the query.

The model choice and model building depends on the purpose of the AI system. Key properties of AI models, such as the degree of transparency and/or explainability (Principle 1.3), the level of robustness (Principle 1.4), and human rights, privacy and fairness implications (Principle 1.2), depend on the type of model, and the model building and inferencing processes (see possible questions in Box 3).

It is important to note that AI models can be built to achieve a specific set of objectives but then used with different objectives, as in the case of transfer learning for example (Principle 1.4). Some leading experts stress the importance of very carefully specifying AI objectives to be fulfillment of human goals rather than intelligence and efficiency (Russell, 2019_[17]).

Box 3. Transparency and explainability (Principle 1.3) and safety, security and robustness (Principle 1.4) throughout the AI system's lifecycle

Possible questions to help determine AI system transparency and explainability (Principle 1.3) include:

- Is it clear what the objectives of the system are, *i.e.* is it possible to formalise the problem that the AI system is being asked to solve?
- Does the system provide useful information for understanding its performance and outputs/decisions during deployment?
- Can the determinant data or knowledge a system uses to make decisions be identified?
- Can we verify that two similar-looking cases result in similar outcomes, *i.e.*, assess the consistency and integrity of AI system outcomes?

Possible questions for policymakers to help determine the safety, security, and robustness of AI systems include:

- Do safety metrics exist which can evaluate the safety of a system for a given use case?
- How does the entity deploying the AI system test for safety during development?
- What measures has the entity deploying the AI system taken to do an adversarial evaluation – that is, explore the AI system through the lens of being a 'bad actor' and trying to break it?
- Does the system change significantly if it is trained with variations of the data available?
- Can the AI system in question be formally verified?

A. AI model characteristics

AI model type [core criteria]

- *Symbolic AI models* (i.e., AI models that use human-generated representations): Symbolic AI uses logical representations to deduce a conclusion from a set of constraints. They include rules, ontologies, and search algorithms and rely on explicit descriptions of variables – agents like humans, entities like factories, objects like machines, variables that can be stock conditions – and descriptions of the inter-relations between these variables. Symbolic models are expressed in languages such as mathematical logic (if-then statements or more abstract ways of representing knowledge via mathematical formulae), agent-based models, event driven models, etc (see example in Box 4). Symbolic AI is still in widespread use for optimisation and planning tools.
- *Statistical AI models* (e.g., genetic algorithms, neural networks, and deep learning): Statistical AI models identify patterns based on data rather than expert knowledge. They have seen increasing uptake recently. These statistical models were previously used primarily for recognition purposes (for instance, translating writing on cheques into machine-readable code), and more recently are also used for tasks like generation (for instance, systems that can synthesise and generate images, or audio). Models that rely on data are designed to be able to effectively extract and represent knowledge from data rather than to contain explicit knowledge.
- *Hybrid AI models*: A number of applications combine symbolic and statistical AI models, i.e. are “hybrid” models. For example, natural language processing algorithms often combine statistical approaches building on large amounts of data and symbolic approaches that consider issues such as grammar rules.

Box 4. Example of a symbolic AI model

A symbolic AI model can be used to represent a car engine production system composed of different factories. Each factory may have different machines assembling parts, operated by teams with different skills working on specific shifts. Parts can be sent from one factory to another using different logistics systems. The specificities and processing rules of this heterogeneous system (human, factories, machines, stock, finances, etc.) are codified based on expert knowledge, describing each specific part of the system and its interactions with the rest. The AI model is built, validated and calibrated based on this knowledge and can be used to simulate possible demand in order to optimise production mechanisms in response to demand volatility.

An AI model’s degree of explainability is primarily determined by the design of the AI model and is linked to the complexity of the system. The more complex the model, the harder it is to explain. Explainability means enabling people affected by the outcome of an AI system to understand how the outcome was arrived at.

Sub-models of rules-based (symbolic) models can often be understood, making it comparatively straightforward to find certain types of errors. By contrast, certain types of machine-learning systems, notably neural networks, are abstract mathematical relationships between factors that can be impossible for humans to understand.

Hybrid AI that combine models built on both data and human expertise is viewed as promising to help address the limitations of both approaches. Hybrid models can provide visibility on complex situations or environments with many interactions, and help to predict what may happen in the future, to help inclusive and sustainable growth and well-being (Principle 1.1).

Discriminative and generative models [optional criteria]

- *Discriminative (or “conditional”) model*: focuses on predicting data labels by learning to distinguish between dataset classes. They are more robust to outliers, but cannot generate new data and can misclassify data points. Examples of discriminative models include linear or logistic regression, support vector machine (SVM), traditional neural networks, decision trees and random forests.
- *Generative model*: focuses on explaining how the data was generated by learning the distribution of dataset classes. These models are capable of generating new data and often perform better than discriminative models on smaller datasets. However, the presence of outliers significantly affects these models. Examples of generative models include naïve bayes models, hidden markov models, linear discriminant analysis (LDA) and generative adversarial networks (GANs).

For policy, whether a model is discriminative or generative determines the type of output that it generates: outputs from discriminative models are predictions whereas outputs from generative models are artefacts.

B. Model building

Model building from machine-learned or human-encoded knowledge

48. The model building process is often called “training” or “optimisation”. Objectives (*e.g.*, output variables) and performance measures (*e.g.*, accuracy, resources for training, and representativeness of the dataset) guide the model building process. AI systems using machine learning for model building have seen tremendous uptake over the past few years. Machine learning is a set of techniques to allow machines to learn in an automated manner through patterns and inferences rather than through explicit instructions from a human. Machine learning approaches often teach machines to reach an outcome by showing them many examples of correct outcomes. However, they can also define a set of rules and let the machine learn by trial and error. Machine learning contains numerous techniques that have been used for decades and range from linear and logistic regressions, decision trees and principle component analysis to deep neural networks. A model can be built from data that is labelled or unlabelled, resulting in different machine learning paradigms: whether data is labelled or not may already be inferred by the section on Format of data and metadata [optional criteria]. When analysing AI systems, systems can be roughly grouped into how much emphasis they place on human-encoded knowledge, versus machine-learned knowledge:

- *Acquisition from human-written rules* (for example writing rules): Human-written rules capture relationships between elements of the environment by logical rules let an AI model deduce a conclusion from a set of constraints and data. They require that researchers build detailed and human-understandable decision structures to translate real-world complexity and help machines make decisions.
- *Acquisition from data through supervised learning*: AI models identify a relationship between input dimensions and labelled target dimensions.
- *Acquisition from data through unsupervised learning*: AI models identify a relationship between input data points based on their similarity.
- *Acquisition from data through semi-supervised learning*: uses both labelled and unlabelled data to identify a relationship between input dimensions and labelled target dimensions.
- *Acquisition from data, augmented by human-encoded knowledge*: ‘Hybrid systems’ combining human-encoded knowledge with knowledge acquired from data are common: for example, self-driving cars are frequently built using complex human-encoded rulesets that encode laws about how to drive - acceptable turns, speed limits, aspects related to braking speeds and tolerances, and so on. These rulesets are then combined with vision systems typically based on neural networks,

which have acquired their capabilities via supervised learning on datasets annotated by the self-driving car companies.

AI systems are only “as good” as the data they are trained on or expert input they are built on. Machine learning systems can make predictions about data similar to that on which they were trained, from which they derive associations and patterns. Machine learning systems can fail in settings that are meaningfully different from those encountered in training.

Data labeling is the process of a tagging data samples, which generally required human knowledge to build training data. Data labeling is critical and can itself require some explainability in contexts such as content moderation, where assigning a label such as “misinformation” or “violent” is critical. In supervised learning the label itself represents extra knowledge that has most often been provided by ‘a human in the loop’, while in unsupervised learning, a human did not label such content.

Expert systems have their own limitations as they require human to build detailed decision structures to translate real world complexity and help machines produce outputs.

Data interaction in machine learning models that evolve in the field

49. In some cases, machine learning models can continue to evolve / acquire abilities from interacting directly with data in the model building process in the following ways:

- *No evolution during operation (no interaction):* the dataset is static and does not change over time. An AI model is given a dataset and learns patterns or associations from it.
- *Evolution during operation through active interaction (including uncontrolled learning):* the AI model actively interacts with the environment and receives data based on these interactions. An example of such a setting is a robot arm, which learns to perform a task (for example, to pick up a cup) by repeatedly attempting to perform the task and receiving feedback on which movements were successful and which movements were not.
- *Evolution during operation through passive interaction:* the AI model receives a continuous stream of data (for example, stock prices), which the system is unable to affect but to which it needs to adapt.



The degree to which the model evolves in response to data is particularly relevant to public policy for AI systems that can iterate and evolve over time and may change their behaviour in unforeseen ways ‘in the field’. Model evolution and model drift are directly relevant to safety, security and robustness (Principle 1.4) as well as accountability and liability (Principle 1.5).

AI models using static data are comparatively more stable. There may be a trade-off between the adaptive nature of an AI system (i.e. whether the model evolves in the field based on input from its environment) and the quality of its outcomes. This trade-off may be more acute with real-time data, as more conflicting data may arrive faster, compromising the quality of the outcomes (Principle 1.4).

Central or federated learning [optional criteria]

50. Models can be trained centrally or in a number of local servers or ‘edge’ devices:

- *Centralised learning:* centralised learning uploads all datasets to a central processing environment to train an algorithm. All datasets are considered local to the training environment. Most current machine learning is centralised.
- *Federated learning:* federated or collaborative learning trains an algorithm across multiple processing environments, which can include edge devices or simply different data centers. Data

samples are kept locally within each environment and not copied across environments. There is no centralised complete dataset for the algorithm to train on.¹⁶

Federated learning helps to address critical issues like privacy, data security and data access rights by building models without sharing data. It also distributes the computing requirements to train an AI system, but may increase the latency.

C. Model inference, i.e. using a model [optional criteria]

51. The same AI model can be used in many different ways. Inference is the process of using an AI model – trained from data or manually encoded – to derive a prediction, recommendation or other outcome based on “new” data that the model was not trained on. Different inference “strategies” can be used to derive varying results from the same model. These strategies are usually designed to optimise specific objectives and performance measures like robustness, accuracy, speed, business metrics or other criteria.

52. For instance, deterministic inference can be applied when random variation is not considered in the model and its outcomes can be fully determined by the parameter values. Reasoning techniques used in expert systems are an example of deterministic inference. If random variation is a major component of the context in which the model operates, then probabilistic inference may be more appropriate. Different methods exist to compare different inference strategies.

AI system explainability (Principle 1.3) relates to the use of the model and to how easily and thoroughly the structure and outputs of an AI system can be understood, and by whom, i.e., understanding the link between input and output.

Deterministic and probabilistic models

- *Deterministic models:* follow precise rules (“if this, then that”) and generate one single outcome.
- *Probabilistic models:* infer several possible models to explain data: deciding which model to use is uncertain. The outcomes made using these different models are also uncertain. Probabilistic models quantify these uncertainties.¹⁷ The different outcomes are associated with different levels of, for instance, performance measures like level of confidence, robustness or risk that can be optimised with different inferencing techniques.¹⁸
- *AI systems can combine both deterministic and probabilistic models.*

For policy, whether a model is probabilistic is relevant to testing and testability (Principle 1.4) as well as to explainability (Principle 1.3). Probabilistic models can generate multiple outcomes with information about their uncertainty. Given the randomness element in probabilistic models, a specific outcome may not easily be reproducible (Principle 1.3 and 1.4).

Key AI actors in the “AI model” dimension: developers and modellers

53. Model building and interpretation involves the creation or selection of models/algorithms, their calibration and/or training and inferencing (i.e. use). It also involved verification and validation whereby models are executed and tuned, with tests to assess performance across various dimensions and considerations. Model building and inferencing involves expertise such as modellers, model engineers, data scientists, domain experts. Model verification and validation currently involves expertise such as data scientists, data/model/systems engineers, governance experts. Actions performed by developers and modellers include:

- *Verification and validation:* to execute and tune models, including metrics to authorise the system for broader deployment.
- *Testing to assess performance:* across various dimensions and considerations.

4) TASK AND OUTPUT

54. The task and output dimension describes what the AI system does: the task it performs and the action that derives from it. The AI system’s model produces recommendations, predictions or other outcomes for a given set of objectives. A human or a machine (an “actuator”) acts upon those recommendations, predictions or outcomes to influence the environment in which the AI system operates, at varying levels of autonomy.

A. Task of the system [core criteria]

55. The task of an AI system refers to what it does, *i.e.* the function that it performs.¹⁹ The following seven categories cover most tasks performed by AI systems (Table 4):





- *Recognition*: identifying and categorising data (e.g., image, video, audio, and text) into specific classifications 
- *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.

Table 4. Tasks of an AI system

	What it does	Type of learning/ reasoning	Examples
Recognition	Identifies and categorises data (<i>e.g.</i> , image, video, audio, and text) into specific classifications. The output is often one label, <i>e.g.</i> , ‘this is a cat’	Supervised classification	Image & object detection; facial recognition; audio, sound, handwriting & text recognition; gesture detection
Event detection	Connects data points to detect patterns as well as outliers or anomalies.	Uses non machine learning cognitive approaches as well as machine learning. Event detection increasingly uses unsupervised and reinforcement learning techniques in which the AI system does not know what it is looking for.	Fraud & risk detection, flagging human mistakes, intelligent monitoring
Forecasting	Uses past and existing behaviours to predict future outcomes, generally to help make decisions. Contains a clear temporal dimension.	Forecasting tends to use machine learning techniques – such as supervised learning - is adaptive and helps improve forecasting over time. Forecasting is generally used for decision support. It may include descriptive analytics; predictive analytics; and projective analytics.	Assisted search, predicting future values for data, predicting failure, predicting population behaviour, identifying and selecting best fit, identifying matches in data, optimising activities, intelligent navigation. 
Personalisation	Develops a profile of an individual, and then learns and adapts to that individual over time. The output is usually a ranking, <i>e.g.</i> , a search engine ranking		Recommender systems based on search & browsing (Netflix, Amazon), personalised fitness, wellness, finance.
Interaction support	Interprets and creates content to power conversational and other interactions between machines and humans (<i>e.g.</i> , involving voice, text, images). Can be real-time or not.	Interaction tends to use semi-supervised learning, enabling models to evolve.	Chatbots, voice assistants, sentiments model, and intent analysis.

Goal-driven optimisation	Gives systems a goal and the ability to find the optimal solution to a problem, which can be by learning through trial and error. It assumes a cost function is given.	Goal-driven optimisation is not necessarily a number. It can be called prescription when based on an optimisation.	Game playing, resource/logistics optimisation, iterative problem solving, bidding and advertising, real-time auctions, scenario simulation.
Reasoning with knowledge structures	Infers new outcomes that are possible even if they are not present in existing data, through modelling and simulation.	This task involves causal reasoning rather than correlation and uses AI techniques beyond machine learning.	Expert systems, legal argumentation, medical diagnosis, planning 

A few policy considerations associated with the tasks performed by AI systems include:

Recognition systems require data that is representative and unbiased to function appropriately. Recognition of people and biometrics, such as facial recognition or voice recognition systems can raise concerns in relation to human rights (Principle 1.2) and robustness and security in case of adversarial attacks (Principle 1.4).

Event detection can benefit people and planet (Principle 1.1), safety and security (Principle 1.4) yet in some contexts raise human rights concerns (Principle 1.2) when used to monitor individuals' activity.

In **forecasting** depending on the application, keeping a human in the loop may be important for accountability (Principle 1.5).

Personalisation can impact social structures and well-being both positively (1.1 benefits) but can also conflict with human values and individuals' right to self-determination (Principle 1.2.) as it tends to provide people with content that they have liked before or that similar people have liked; contributing to disinformation and echo chamber effects.

Interaction support tasks in which an AI system interacts with people may implicate data usage and data privacy (Principle 1.2) and may require higher transparency and disclosure of the fact that one is interacting with a chatbot (Principle 1.3). It may also impact labour markets (Principle 2.4).

Goal-driven optimisation and other similar tasks using machine learning algorithms that can learn from themselves through trial and error and may require humans in or on the loop (Principle 1.5). Limits on the power of this type of system may be needed if exponential growth occurs (e.g., artificial general intelligence).

Reasoning with knowledge structures is promising to help inclusive and sustainable growth and well-being (Principle 1.1) by allowing for the simulation of different scenarios considering causal and counterfactual relationships and situations that change with time, such as improving legacy power generation systems.²⁰

Action autonomy level [core criteria]

56. A human or a machine (an “actuator”) uses the outcome from the AI system (specifically, the outcome from the inferencing process) to perform an action (prescribed by humans) that influences the environment in which the system operates. The way in which this action is performed determines the autonomy level of an AI system; that is, the degree to which a system can act without human involvement.

57. This section considers four variations in the degree of system autonomy using a typology that originated in the field of aviation (Endsley, 1987_[18]):


- *No action autonomy* (also referred to as “human support”): the AI system cannot act on its recommendations or output. The human uses or disregards the AI system’s recommendations or output at will.
- *Low action autonomy* (also referred to as “human-in-the-loop”): the system evaluates input and acts upon its recommendations or output if the human agrees.

- *Medium action autonomy* (also referred to as “human-on-the-loop”): the system evaluates input and acts upon its recommendations or output unless the human vetoes.
- *High action autonomy* (also referred to as “human-out-of-the-loop”): the system evaluates input and acts upon its recommendations or output without human involvement.

High action autonomy systems pose important policy considerations, in particular when deployed in critical functions and activities or in contexts that may put human rights or fundamental values (Principle 1.2) at risk.

C. Displacement potential

58. The ability of an AI system to automate tasks that had previously been, or are currently being executed by humans is dependent on a variety of task-dependent factors (e.g., perception and manipulation requirements, uncertainty, and creative and social intelligence factors).²¹ Historically, automation has been limited to tasks that require perception and manipulation of homogeneous objects with clearly defined processes and limited uncertainty; are conducted within controlled environments; and require no creativity or social interaction.


59. However, recent innovations in AI are changing this landscape and are starting to include tasks that have been typically executed by higher-skilled workers. For simplicity, an AI system’s automation potential can be split into two categories 

- *High displacement potential*: AI systems execute tasks similar to those that require clearly defined processes with well understood outputs (e.g., tasks performed by clinical lab technicians, optometrists, chemical engineers, actuaries, credit analysts, accountants, operations research analysts, concierges, mechanical drafters, brokerage clerks, and quality control inspectors).
- *Low displacement potential*: AI systems execute tasks similar to those that require reasoning about novel situations (e.g., research), occupations requiring interpersonal skills (e.g., teachers and managers, some baristas) and physical occupations that require perception and manipulation of a plurality of irregular objects in uncontrolled environments with limited room for mobility (e.g., maids, cleaners, cafeteria attendants, hotel porters, roofers and painters, massage therapists, and plasterers and stucco masons).

An AI system’s capacity to automate tasks can have an impact on the world of work (Principle 2.4). It should be noted that “high displacement potential” does not imply a high likelihood of being replaced by AI, as this would require a more complex assessment of the technical feasibility and context of the task. Likewise, “low displacement potential” does not mean that the occupation will not see significant automation of key tasks.

D. Combining tasks and actions into composite systems [optional criteria]

60. AI systems frequently perform several tasks before producing an outcome that influences the environment. Several composite systems that combine different tasks are common and constitute well-known areas and communities. They may generate specific policy considerations that differ from those produced by single-task systems.

61. The following are examples of common AI systems or application areas that combine several tasks and, in some cases, also actions: 

- *Content generation* (also referred to as synthesis): content generation includes generating new images, video, text, and audio. This task combines forecasting and recognition tasks. However, the output often combines several existing elements such as images, text and audio to produce an object that was never seen before. This task tends to use structured learning. Examples of content

generation include machine translation, generative art, news stories – including fake news – spam emails and “deep fake” videos.

- *Autonomous systems*: robotics increasingly embeds different tasks – such as recognition and goal-driven optimisation – to perform an action in the real world. Similarly, autonomous systems perform a recognition task, based on which they try to find an optimal path to arrive at the best solution, and then act accordingly. In an autonomous vehicle, this could be a recommendation to turn left or right to minimise travel time, followed by the execution of the action. This system is autonomous in the sense that it does not require human supervision to act on its environment, but the way in which it performs its task (recognition) is not autonomous, as it relies on supervised learning.
- *Control systems*: control systems manage, command, direct, or regulate the behaviour of other devices or systems using control loops. Control systems generally assess environments through recognition, event detection or forecasting and propose a goal-driven action. They range from domestic heating controllers to large industrial control systems that use reinforcement learning to manage processes or machines. Control systems are common in the context of robotics or factories.

From a policy perspective, content generation has relevant implications for human rights and democratic values (Principle 1.2), particularly when the content generated is realistic enough to be confused with “real” content. AI content generation amplifies the need to provide meaningful information to make stakeholders aware of their interactions with AI systems, to enable those affected to understand and to challenge the outcome (Principle 1.3; Principle 1.5). AI content generation also raises intellectual property right questions (e.g. on the patentability of AI-assisted inventions and copyrighting of AI-generated creative work).

Autonomous systems and control systems have received increased attention and have direct implications for human safety (Principle 1.4) and accountability (Principle 1.5).

E. Core application areas [core criteria]

- *Human language technologies*: language technologies analyse, modify, produce or respond to human text and speech. Human language technologies may combine tasks like recognition, personalisation and interaction support.
- *Computer vision*: computer vision is concerned with training computers to interpret and understand the visual world. Feeding digital images and videos into deep learning models, machines can identify and classify objects and react to what they “see.” Computer vision may include tasks like object recognition and event detection.
- *Robotics*: robotics increasingly embeds different tasks – such as recognition and goal-driven optimisation – to perform an action in the real world. Autonomous systems similarly perform recognition tasks, based on which they try to find an optimal path to arrive at the best solution, and then act accordingly. In an autonomous vehicle, this could be a recommendation to turn left or right to minimise travel time, followed by the execution of the action.

Key AI actors in the “Task and output” dimension: system integrator

62. The Task and output dimension can be associated with the ‘deployment’ stage of the AI system lifecycle (OECD, 2019[7]). Deployment into live production involves piloting, checking compatibility with legacy systems, ensuring regulatory compliance, managing organisational change, and evaluating user experience. Deployment currently involves expertise such as system integrators, developers, systems/software engineers, testers and domain experts.

II. APPLYING THE FRAMEWORK

63. The framework can be used to assess applications (such as a credit scoring application, AlphaGo Zero or a hybrid system to manage a manufacturing plan, examples 1, 2 and 3). It can also be used to some extent to assess some general characteristics of tools such as GPT-3 (example 4), although much of the functionality will depend of the final application of the tool.

Example 1: A credit-scoring system

A credit-scoring system illustrates a machine-based system that influences its environment (whether people are granted a loan). It makes recommendations (a credit score) for a given set of objectives (credit-worthiness). It does so by using both machine-based inputs (historical data on people's profiles and on whether they repaid loans) and human-based inputs (a set of rules). With these two sets of inputs, the system perceives real environments (whether people have repaid loans in the past or whether they are repaying loans on an ongoing basis). It abstracts such perceptions into models automatically. A credit-scoring algorithm could, for example, use a statistical model. Finally, it uses model inference (the credit-scoring algorithm) to formulate a recommendation (a credit score) of options for outcomes (providing or denying a loan).

A. Context

- *Sector:* Section K (Financial and insurance activities)
- *Business function:* Sales, customer service
- *Critical functions:* Yes
- *Scale of deployment and technology maturity:*
 - *Scale of deployment:* Broad
 - *AI system maturity:* Actual system or subsystem in final form in operational environment – TRL 9
- *User:* Amateur (bank employee)
- *Impact:*
 - *Impacted stakeholders:* Consumers
 - *Optionality:* Not optional / cannot opt out
 - *Business model:* For-profit use – other model
- *Benefits and risks to human rights and democracy*

AI outcomes impact human rights or democratic values:	No impact	Outcome dependent
Human dignity, life and physical and mental integrity	X	
Liberty and security	X	
Fair trial; no punishment without law; effective remedy		X
Privacy and family life, and the protection of personal data		X
Freedom of thought, conscience and religion	X	
Freedom of expression; assembly and association	X	
Non-discrimination and equal treatment		X
Protection of property and peaceful enjoyment of possessions		X
Right to education		X
Right to democracy and free elections	X	
Rights of the child, the elderly and persons with disabilities		X
Other (detail)	X	

- *Benefits and risks to well-being*

AI outcomes impact well-being:	No impact	Outcome dependent
Health (including mental health)	x	
Housing		x
Income and wealth		x
Work and job quality	x	
Environment quality	x	
Social connections	x	
Civic engagement	x	
Education	x	
Subjective well-being	x	
Work-life balance	x	

B. Data and input

- *Provenance, collection and dynamic nature*
 - *Collection:* Collected by humans (set of rules) and automated sensing devices (*e.g.*, loan payments)
 - *Provenance:* provided by experts (rules), provided by loan candidate (*e.g.*, personal information), observed by the algorithm (*e.g.*, history of payments), derived data (*e.g.*, credit rating and other scores)
 - *Dynamic nature:* static (*e.g.*, gender); dynamic data updated from time to time (*e.g.*, salary) and dynamic real-time data (*e.g.*, day-to-day payments)
 - *Scale:* TBD (small or medium)
- *Structure and format*
 - *Structure:* Structured data
 - *Data and metadata format:* Standardised
- *Rights and ‘identifiability’:*
 - *Rights:* Personal and proprietary
 - *Type of personal data:* Identified data
- *Appropriateness and quality:* quality unknown; appropriate data

C. AI model

- *AI model characteristics:*
 - *Model type:* Hybrid
 - *Discriminative vs generative:* Discriminative (a score can be seen as a probability)
- *Model building:*
 - *Model building:* Acquisition from data, augmented by human-encoded knowledge
 - *Central vs distributed:* Central
 - *Data interaction:* Evolution during operation through passive interaction
- *Model inference:*
 - *Deterministic vs probabilistic:* Deterministic

D. Task and output

- *Task:* Forecasting, reasoning with knowledge structures
- *Action autonomy:* Low
- *Displacement potential:* Medium TBD
- *Composite system:* Yes
- *Core application area:* TBD

Example 2: AlphaGo Zero

AlphaGo Zero is an AI system that plays the board game Go better than any professional human Go players. The board game's environment is virtual and fully observable. Game positions are constrained by the objectives and the rules of the game. AlphaGo Zero is a system that uses both human-based inputs (the rules of the game of Go) and machine-based inputs (learning based on playing iteratively against itself, starting from completely random play). It abstracts the data into a (stochastic) model of actions ("moves" in the game) trained via so-called reinforcement learning. Finally, it uses the model to propose a new move based on the state of play.

A. Context

- *Sector:* Section R (Arts, Entertainment, and Recreation)
- *Business function:* N/A
- *Critical functions:* No
- *Scale of deployment and technology maturity:*
 - *Breadth of deployment:* Narrow
 - *AI system maturity:* System or subsystem complete and qualified through test and demonstration – TRL 8
- *User:* AI expert practitioner (e.g., DeepMind engineers)
- *Impact:*

- *Impacted stakeholders*: None for now. If deployed in production, specific communities (e.g., Go players)
- *Optionality*: N/A
- *Business model*: Non-profit use (outside public sector)
- *Benefits and risks to human rights and democracy*

● AI outcomes impact human rights or democratic values:	No impact	Outcome dependent
Human dignity, life and physical and mental integrity	X	
Liberty and security	X	
Fair trial; no punishment without law; effective remedy	X	
Privacy and family life, and the protection of personal data	X	
Freedom of thought, conscience and religion	X	
Freedom of expression; assembly and association	X	
Non-discrimination and equal treatment	X	
Protection of property and peaceful enjoyment of possessions	X	
Right to education	X	
Right to democracy and free elections	X	
Rights of the child, the elderly and persons with disabilities	X	
Other (detail)	X	

- *Benefits and risks to well-being*

AI outcomes impact well-being:	No impact	Usage dependent
Health (including mental health)	X	
Housing	X	
Income and wealth	X	
Work and job quality	X	
Environment quality	X	
Social connections	X	
Civic engagement	X	
Education	X	
Subjective well-being	X	
Work-life balance	X	

B. Data and input

- *Provenance, collection and dynamic nature*
 - *Collection*: Collected by humans (the rules of the game of Go) and automated sensing devices
 - *Provenance*: provided by experts (the rules of the game of Go), observed by the algorithm, and synthetic data
 - *Dynamic nature*: static (human knowledge) and dynamic real-time data (each move in the game)
 - *Data scale*: TBD (large or very large)
- *Structure and format*
 - *Structure*: Complex structured data
 - *Data and metadata format*: standardised and non-standardised

- *Rights*: Public and proprietary
- *Appropriateness and quality*: Representative and appropriate, low noise/missing values/outliers

C. AI model

- *AI model characteristics*:
 - *Model type*: Hybrid
 - *Discriminative vs generative*: Generative
- *Model building*:
 - *Model building*: Acquisition from data, augmented by human-encoded knowledge
 - *Data interaction*: Evolution during operation through active interaction
 - *Central vs distributive*: Central
- *Model inference*:
 - *Deterministic vs probabilistic*: Both

D. Task and output

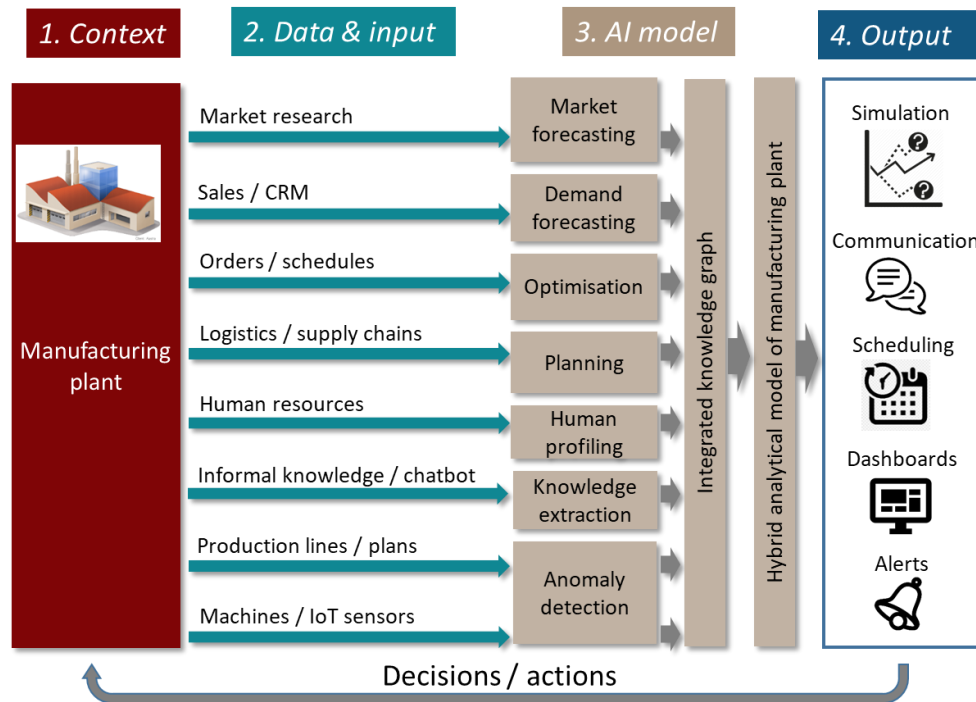
- *Task*: Forecasting, goal-driven optimization, reasoning with knowledge structures
- *Action autonomy*: High
- *Displacement potential*: Low TBD
- *Composite system*: Yes
- *Core application area*: Human language technologies and/or computer vision (TBC)

Example 3: Qlector.com LEAP system to manage a manufacturing plant

An AI system controlling and running a manufacturing plant provides an example of a complex hybrid AI system. The context is the physical manufacturing plant factory floor. Different AI models, associated with different data sources, perform particular activities for the factory based in different data and input. These modelling activities include: event detection (anomaly detection based on data from machines in production lines); goal-driven optimisation (based on orders and schedules, logistics and supply chain data); reasoning with knowledge structures (simulations); interaction support (customer relationship management); demand forecasting (based on sales) and strategic market forecasting (based on market research).

All these models are combined into a large evolving knowledge graph with a symbolic AI type of data structure that interconnects the different levels of the factory (Figure 8). The resulting model is a hybrid analytical model of a manufacturing plant that some could refer to as a digital twin of the factory. The outputs of the model include: alerts, information on dashboards, scheduling, communications with customers, and simulations of possible futures to inform decisions. While humans are often involved in the actions resulting from the system outputs, factory processes are increasingly autonomous. The output/decision feeds back into the context/physical environment.

Figure 7. AI system to help manage a manufacturing plant



Source: Example taken from Qlector.com LEAP system

A. Context

- *Sector:* Section C (Manufacturing)
- *Business function:* many (including sales; customer service; planning and budgeting; procurement; logistics; human resource management; monitoring and quality control; production; maintenance)
- *Critical functions:* No
- *Scale of deployment and technology maturity:*
 - *AI system maturity:* actual system or subsystem in final form in operational environment – TRL 9
 - *Breadth of deployment:* Narrow
- *Impact:*
 - *Impacted stakeholders:* Consumers, Workers / employees, Business
 - *Optionality:* variable
 - *Business model:* For-profit use
- *User:* Amateurs, non-expert and expert practitioners
- *Benefits and risks to human rights and democracy*

AI outcomes impact human rights or democratic values:	No impact	Outcome dependent
Human dignity, life and physical and mental integrity		X
Liberty and security	X	
Fair trial; no punishment without law; effective remedy	X	
Privacy and family life, and the protection of personal data		X
Freedom of thought, conscience and religion	X	
Freedom of expression; assembly and association	X	
Non-discrimination and equal treatment		X
Protection of property and peaceful enjoyment of possessions	X	
Right to democracy and free elections	X	
Rights of the child, the elderly and persons with disabilities	X	
Right to education	X	
Other (detail)	X	

- *Benefits and risks to well-being*

AI outcomes impact well-being:	No impact	Usage dependent
Health (including mental health)		X
Housing	X	
Income and wealth		X
Work and job quality		X
Environment quality	X	
Social connections		X
Civic engagement	X	
Education	X	
Subjective well-being		X
Work-life balance		X

B. Data and input

- *Provenance, collection and dynamic nature*
 - *Collection*: Collected by humans and by automated sensing devices
 - *Provenance*: all (expert input, provided data, observed data, synthetic data and derived data)
 - *Dynamic nature*: static (human knowledge) and dynamic real-time data (data from machines in production lines)
 - *Data scale*: medium
- *Structure and format*
 - *Structure*: all (unstructured, semi-structured, unstructured, complex structured data)
 - *Data and metadata format*: standardised and non-standardised
- *Rights*: proprietary
- *Appropriateness and quality*: Representative and appropriate, noise /missing values/outliers

C. AI model

- *AI model characteristics:*
 - *Model type:* Hybrid
 - *Discriminative vs generative:* discriminative and generative
- *Model building:*
 - *Model building:* Acquisition from data, augmented by human-encoded knowledge
 - *Data interaction:* Evolution during operation through active and passive interaction
 - *Central vs distributive:* federated
- *Model inference:*
 - *Deterministic vs probabilistic:* Both

D. Task and output

- *Task:* all (recognition, event detection, forecasting, interaction support, goal-driven optimization, reasoning with knowledge structures)
- *Action autonomy:* Medium action autonomy
- *Displacement potential:* High
- *Composite system:* Yes
- *Core application area:* Human language technologies (as well as process planning and optimisation and Internet-of-things)

Example 4: GPT-3

GPT-3 is a large, pre-trained language model that has the capacity to search over, generate, and manipulate strings of text. GPT-3 can take in arbitrary inputs in the form of text strings, which lead to it generating an output. GPT-3 can be conditioned with up to 2048 distinct characters, letting it learn from the examples it is primed with.

GPT-3 is a general purpose AI system, meaning it can theoretically be used to deploy applications in any sector of the economy. Such applications would need to be considered within their specific socio-economic context; for example, a creative writing application built with GPT3 should be treated differently to one that seeks to give a user medical advice in response to a query. Examples of the use of GPT3 include text classification activities to search over news articles, and generation of emails from a summary sentence. The earlier example is run through the classification framework below.

A. Context

- *Sector:* Section J (Information and Communication)
- *Business function:* Any
- *Critical functions:* No
- *Scale of deployment and technology maturity:*

- *AI system maturity*: System prototype demonstration in operational environment – TRL 7
- *Breadth of deployment*: Narrow
- *Use*: For-profit use – other model (e.g., business intelligence) or non-profit use (e.g., research, journalism)
- *Impact*:
 - *Impacted stakeholders*: Workers (e.g., could lead to automation of some tasks)
 - *Optionality*: Optional / can opt out
- *User*: Amateur
- *Benefits and risks to human rights and democracy*

AI outcomes impact human rights or democratic values:	No impact	Outcome dependent
Human dignity, life and physical and mental integrity	X	
Liberty and security	X	
Fair trial; no punishment without law; effective remedy		X
Privacy and family life, and the protection of personal data	X	
Freedom of thought, conscience and religion		X
Freedom of expression; assembly and association	X	
Non-discrimination and equal treatment		X
Protection of property and peaceful enjoyment of possessions	X	
Right to democracy and free elections		X
Rights of the child, the elderly and persons with disabilities	X	
Right to education	X	
Other (detail)	X	

- *Benefits and risks to well-being*

AI outcomes impact well-being:	No impact	Outcome dependent
Health (including mental health)	X	
Housing	X	
Income and wealth	X	
Work and job quality		X
Environment quality	X	
Social connections	X	
Civic engagement	X	
Education		X
Subjective well-being	X	
Work-life balance	X	

B. Data and input

- *Provenance, collection and dynamic nature*
 - *Provenance*: Observed and derived
 - *Collection*: Collected by humans and automated sensing devices (e.g., collected by humans with subsequent filtering by machines and humans)
 - *Dynamic nature*: dynamic data updated from time to time
 - *Scale*: very large

- *Structure and format*
 - *Structure*: Unstructured data
 - *Data and metadata format*: Non-standardised
- *Rights and ‘identifiability’*: Public and proprietary
- *Appropriateness and quality* : Noisy data that is, by design, highly representative and diverse with regard to a large part of (predominantly English) text and code found on the internet; appropriate data

C. AI model


- *AI model characteristics*:
 - *Model type*: Statistical (data-driven)
 - *Discriminative vs generative*: Generative
- *Model building*:
 - *Model building*: Acquisition from data, augmented by human-encoded knowledge
 - *Central vs distributive*: Central
 - *Data interaction*: Evolution during operation through passive interaction
- *Model inference*:
 - *Deterministic vs probabilistic*: Deterministic


D. Task and output

- *Task*: Reasoning with knowledge structures; interaction support; recognition; personalisation
 - *Action autonomy*: Low
 - *Displacement potential*: TBD
 - *Composite system*: Yes
 - *Core application area*: Human language technologies
- OpenAI also published a [‘Model Card’ about GPT-3](#).

III. ILLUSTRATIVE ETHICAL AND SOCIETAL RISK ASSESSMENT BASED ON THE FRAMEWORK

Overview and approach

64. Policy makers favour a risk-based approach to regulating AI in order to focus oversight and intervention where it is most needed while avoiding unnecessary hurdles to innovation. The OECD AI Principles state that “AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias” 

65. The risk of using AI strongly depends on the application. Since it is difficult to anticipate and assess every possible use case, AI applications need to be grouped into a small number of risk levels. In its 2019 white paper, the European Commission proposed as the simplest approach distinguishing two risk levels (low/high). Various academic groups as well as expert panels (e.g., the German Data Ethics Commission and IEC SEG10) have proposed four to five risk levels . Typical criteria for determining which risk level an AI application belongs in are the:

- Scale, *i.e.* seriousness of adverse impacts (and probability)
- Scope, *i.e.* breadth of application, *e.g.*, the number of individuals that are or will be affected
- Optionality, *i.e.* the degree of choice as to whether to be subject to the effects of an AI system.

66. It should be noted that these discussions about handling AI risks are complementary to existing well-established risk assessment frameworks e.g., for functional and product safety (OECD, 2016⁽¹¹⁾) or digital security (OECD, 2015⁽¹¹⁾), and hence tend to focus on ethical and societal risks. Existing human rights and responsible business impact assessments guidelines are also relevant (OECD, 2019⁽¹¹⁾).

67. The classification framework presented in this report already describes most aspects of an AI application, including its context, data and input, the AI model, and task and output. Once all the information under this classification has been collected for a specific use case, the framework can also be used for assessing the associated ethical and societal risk.

68. In other words, the classification framework can be mapped to a basic risk assessment without requiring additional input. This has considerable practical significance.

69. The focus in this first phase is on distinguishing the lowest risk level from any higher risk levels. This reflects the fact that this is the most important distinction to make in practice for an AI application: whether any non-trivial ethical and/or societal risks are present or not, and whether therefore a more in-depth analysis (beyond the scope of the present report) and regulation and/or other policy interventions are required.

Demonstration of risk mapping

70. To map the classification framework to a risk assessment, two mechanisms are needed:

- *Systems that are not low risk (“showstoppers”)*: Certain categories in the framework have such a significant impact on risk that each one on its own can push an AI application above the lowest risk level. As an example, if the context of an AI system is such that it can cause immediate


physical harm, this AI application cannot be categorised in the lowest risk level, regardless of the other characteristics of this application.

- *Cumulative effect on risk:* Most categories in the framework have a less clear impact on risk. Some characteristics tend to point towards higher risk (e.g., impact on workers), other characteristics tend to point towards a lower risk (e.g., the use of anonymised rather than identified data). Adding up these indications results in a basic cumulative risk score that provides a practical first illustration of how the present framework could be used in practice. Policy decisions can then determine at which maximum risk score a specific application is still considered to be in the lowest risk level and whether/what weighting is needed in which contexts.

71. The following demonstrates how these two mechanisms are applied (Table 5). Some categories in the framework are not deemed to have a clear positive or negative impact on risk and are therefore shown as neutral.

Table 5. Illustrative AI system risk mapping using the framework

AI system characteristics (by dimension)	Cumulative effect on risk	Not low risk
1) CONTEXT		
Industrial sector	↑ or ↓	
Business function	↑ or ↓	
Impacts critical functions / activities		
AI system is in a critical sector or infrastructure	↑	
AI system performs a critical function independent from its sector	↑	X
Breadth of deployment		
A pilot project	↓	
Narrow deployment (e.g. one company in one country)	↑ or ↓	
Broad deployment (e.g. one sector)	↑	
Widespread deployment (e.g. across countries and sectors)	↑	
AI system maturity		
TRL 1 to 3	↑	
TRL 4 to 7	↑ or ↓	
TRL 8 to 9	↓	
Users of AI system		
Amateur	↑	
Practitioner who is not an AI expert	↑ or ↓	
Practitioner who is an AI expert or system developer:	↑ or ↓	
AI system maturity	↑ or ↓	
Impacted stakeholders		
Consumers	↑	
Workers / employees	↑	
Business	↑ or ↓	
Government agencies / regulators	↑	
Specific communities	↑ or ↓	
Children or other vulnerable or marginalised groups	↑	
Optionality		
Users cannot opt out of using the AI system	↑	
Users can correct or contest AI output	↑ or ↓	
Users can opt-out of using the system	↓	
For-profit use, non-profit use or public sector use		
For-profit use – subscription fee model	↑ or ↓	
For-profit use – advertising model	↑	
For-profit use – other model	↑ or ↓	

Non-profit use (outside public sector)	↑ or ↓	
Public sector use	↑	
Other	↑ or ↓	
Direct and immediate risks of violating human rights or fundamental values (only considering negative impacts)		
Life and physical and mental integrity	↑	X
Liberty and security	↑	X
Fair trial; no punishment without law; effective remedy	↑	X
Privacy and family life	↑	
Freedom of thought, conscience and religion	↑	X
Freedom of expression; assembly and association	↑	X
Non-discrimination	↑	
Protection of property and peaceful enjoyment of possessions	↑	
Right to education	↑	X
Right to democracy and free elections	↑	
Human autonomy	↑	
Human dignity	↑	
Other (detail)	↑	
Direct and immediate risks to individuals' well-being (only considering negative impacts)		
Health (including mental health)	↑	X
Housing	↑	X
Income and wealth	↑	
Work and job quality	↑	
Environment quality	↑	
Social connections	↑	
Civic engagement	↑	
Education	↑	
Subjective well-being	↑	
Work-life balance	↑	

2) DATA AND INPUT	
Provenance of data and input	↑ or ↓
Detection and collection of data and input	↑ or ↓
Dynamic nature of data	
Static data	↓
Dynamic data updated from time-to-time	↑ or ↓
Dynamic real-time data	↑
Scale	↑ or ↓
Structure of data and input	↑ or ↓
Format of data and metadata	
Standardised data format	↑ or ↓
Non-standardised data format	↑
Standardised dataset metadata	↑ or ↓
Non-standardised dataset metadata	↑
Rights associated with data and input	
Proprietary data	↑
Public data	↑ or ↓
Personal data	↑
Identifiability of personal data	
Identified data	↑
Pseudonymised data	↑ or ↓
Unlinked pseudonymised data	↓
Anonymised data	↓

Aggregated data	↓
Data quality and appropriateness	
appropriateness of data for a particular problem	↓
(high) sample representativeness	↓
adequate sample size	↓
(high) completeness and coherence of sample	↓
(low) data noise	↓

3) AI MODEL	
AI model characteristics	
(High) transparency and explainability	↓
(High) safety, security, robustness	↓
(High) reproducibility	↓
Evolution during operation	↑
Evolution through uncontrolled learning	↑
Privacy-preserving properties, e.g. federated learning	↓

4) TASK AND OUTPUT	
Task of the system	
Recognition	↑ or ↓
Event detection	↑ or ↓
Forecasting	↑ or ↓
Personalisation	↑
Interaction support	↑
Goal-driven optimisation	↑ or ↓
Reasoning with knowledge structures	↑ or ↓
Action autonomy level	
High action autonomy	↑
Medium action autonomy	↑
Low action autonomy	↑ or ↓
No autonomy	↓
Displacement potential	
High displacement potential	↑
Core application areas	↑ or ↓


Note: items marked “↑ or ↓” are to be assessed depending on the AI system usage and outcomes.


Next steps

72. The risk assessment approach presented in this section is illustrative. Provided positive feedback on the overall approach, the working group will conduct further analysis on the criteria to include in a risk assessment and how best to aggregate these criteria, noting that different criteria may be inter-dependent. The group will use examples of AI systems in clearly different risk categories to assess the usefulness of different criteria and to try to calibrate in an empirical way if possible. At the same time, it should be noted that there may be a trade-off between developing a simple, user-friendly assessment (which is the goal of the present exercise) and a very accurate assessment as the latter may require significant in-depth information on a model, which may be unknown to the average user.

73. The user-friendly *online survey* that will test the usability of the overall OECD Framework for the Classification of AI Systems during the public consultations could also be used to test the usability of the associated risk assessment framework presented in this section.

Annex A. Sample AI applications by sector, ordered by proxy for diffusion

Sectors	Description	Main applications of AI
Information and communication (Section J)	Includes the production and distribution of information and cultural products, the provision of the means to transmit or distribute these products, as well as data or communications, information technology activities and the processing of data and other information service activities.	Advertising Image or text processing Personalised content generation Augmented and virtual reality Customer services Network security Network management, predictive maintenance Software production
Professional, scientific and technical activities (Section M)	Includes specialised activities that require a high degree of training and make specialised knowledge and skills available to users including legal affairs, management, consultancy, architecture, engineering, R&D, advertising and more.	Legal and accounting AI applications Marketing and advertising services (e.g. personalised advertising and pricing, click prediction systems, recommendations based on social media posts, emails, web navigation etc.) Scientific research and development 
Financial and insurance activities (Sector K)	Includes financial service activities, insurance, reinsurance and pension funding and activities to support financial services, funds and holdings.	Credit scoring Financial technology lending Cost reduction in the front and middle office Fraud detection and legal compliance Insurance Algorithmic trading
Administrative and support service activities (Section N)	Includes a variety of activities that support core business functions and of which the primary purpose is not the transfer of specialised knowledge. This includes security services, renting and leasing, office administrative functions, reservation services.	Auditing expense reports Hiring applications Smart contracts Customer relations
Agriculture, forestry and fishing (Section A)	Includes the exploitation of vegetal and animal natural resources such as growing of crops, raising and breeding of animals, harvesting of timber and other plants, animals or animal products from a farm or their natural habitats.	Agricultural robots and drones Crop and soil monitoring Predictive analytics
Manufacturing (Section C)	Includes the physical or chemical transformation of materials, substances, or components into new product. The materials transformed are products of agriculture, forestry, fishing, mining or quarrying as well as products of other manufacturing activities. Excludes waste.	Market and demand forecasting Product assembly Asset optimisation Supply chain management and planning Anomaly detection
Public administration and defence; compulsory social security (Section O)	Includes activities of a governmental nature, normally carried out by the public administration such as public order and safety, legislative activities, foreign affairs, national defence and more.	Predictive algorithms in the legal system Predictive policing Use of AI by the judiciary Use of AI in defence (e.g. drones footage for surveillance, cyberdefence, command and control, autonomous vehicles)
Wholesale and retail trade (Section G)	Includes wholesale and retail sale (i.e. sale without transformation) of any type of goods and the rendering of services incidental to the sale of these goods. Wholesaling and retailing are the final steps in the distribution of goods.	Customer management (e.g., prediction of customer needs; identification of upsell and cross-sell opportunities; agile response mechanism) Operational efficiency (e.g., just-in-time production/delivery, product categorisation/placement, demand forecasting, check-out free store) Legal efficacy (e.g., compliance systems to predict violations in the supply

	Goods bought and sold are also referred to as merchandise.	chain; legal contract translation, cataloguing and implementation - "smart contracts") Customer acquisition (e.g., matching buyers and sellers, personalised ads/referrals – see "Marketing and advertising services"). Customer retention (e.g., learning and predicting customers' preferences and needs, tailored offers, dynamic pricing) Customer service (e.g., conversational interfaces, voice and video search, chatbots, mood tracking)
Education (Section P)	Includes private and public education, all levels from pre-school to higher education, adult education, sport education, literacy programmes and more.	Personalising learning with AI (e.g. adaptive tests and learning systems) ²² Supporting students with special needs with AI (e.g. wearables using AI) Reducing dropout rates (e.g. predictive and diagnosis models) Chatbots 
Human health and social work activities (Section Q)	Includes the provision of health and social work activities. Activities include a wide range of activities, starting from health care provided by trained medical professionals in hospitals and other facilities, over residential care activities that still involve a degree of health care activities to social work activities without any involvement of health care professionals.	Detection (e.g., outbreak alerts) Precision medicine (e.g., treatments) Optimise health systems (e.g., resource allocation, workflow management) Facilitating health research (e.g., drug discovery, vaccine development) Preventative / personalised healthcare (e.g., self-monitoring tools, applications and trackers) Nursing and elderly care Diagnosis (e.g., radiology)
Transportation and storage (Section H)	Includes the provision of passenger freight transport, whether scheduled or not, by rail, pipeline, road, water or air. Also includes associated activities such as terminal and parking, cargo handling, storage. Includes rental and postal activities.	Warehouse and supply chain management Shipping and itinerary route optimisation, including based on traffic data Autonomous driving systems Computer vision technologies that track driver's eyes / focus to assess distraction
Accommodation and food service activities (Section I)	Includes the provision of short-term stay accommodation for visitors, of complete meals and drinks for immediate consumption. The type of supplementary services provided within this section can vary widely. Excludes long-term stay and primary residence.	AI-powered chatbots (e.g., booking, ordering) Face recognition (check-in) Analysis of customer, occupancy and guest feedback data.
Construction (section F)	Includes general and specialised construction activities for buildings and dwellings, civil engineering works, new work, additions, repairs and alterations.	3D Building Information Modelling (BIM) Buildings simulators Drones and sensors on construction sites Data analytics based on the real-time data collected on-site.

Note: the table is ordered from ISIC REV 4 industry sectors that are seeing the most AI adoption to those that are experiencing the least AI-adoption (see also Annex B).

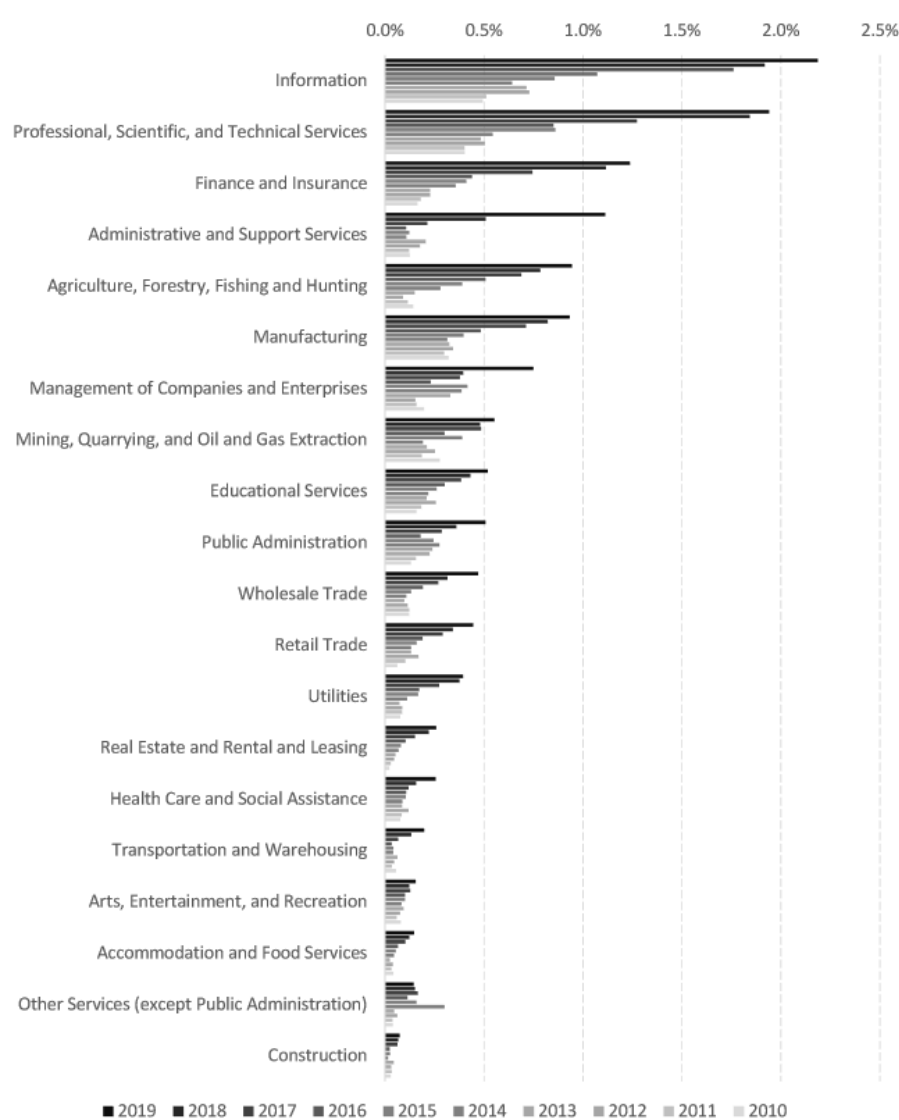
Source: Based on different sources including (OECD, 2019^[2]).

Annex B. AI adoption per industry

74. A number of researchers are using AI labour demand – that is, firms’ jobs data – as a proxy for AI adoption in different industries. Figure 8 shows the percentage of AI-related job vacancies – out of total vacancies – by 2-digit NAICS industries in 2010-2019 (Alekseeva, 2019_[19]).

75. AI is mostly being adopted in industries such as information; professional, scientific and technical services; finance and insurance; administrative and support services; agriculture; management; mining, quarrying, and oil and gas extraction; education; public administration; whole and retail trade; and manufacturing. Adoption of AI in fields such as healthcare or transportation seems to be much lower.

Figure 8. Share of AI jobs by industry (2010-2019)

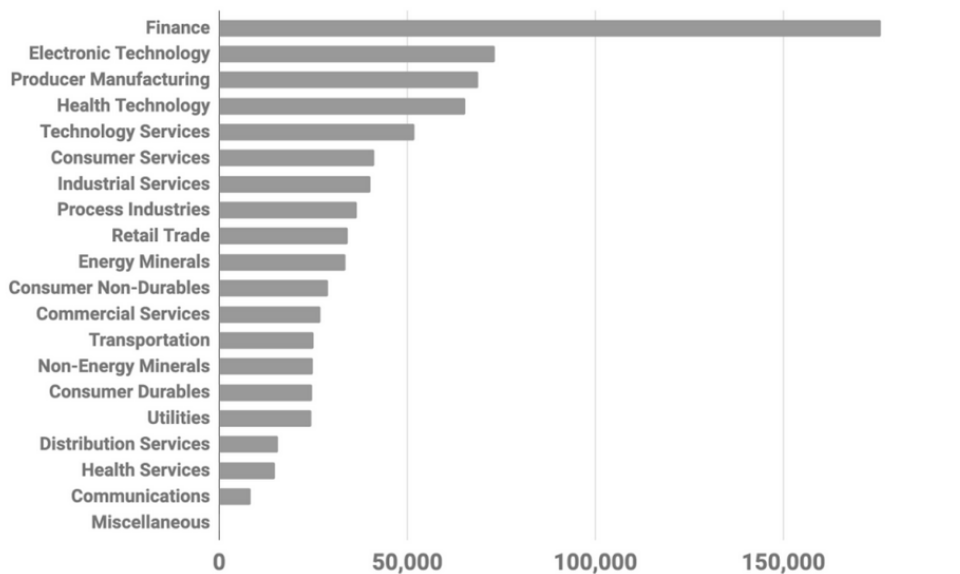


Note: Industries are ranked by the AI share in 2019. Data only includes job postings with non-missing 2-digit NAICS codes.

Source: (Alekseeva, 2019_[19])

76. An alternative measure of AI adoption is mentions of AI in company earnings calls. The share of earnings calls where AI is mentioned has increased substantially in recent years, led by earnings calls in the finance, electronic technology, and producer manufacturing sectors. While mentions of AI are prevalent in the health technology sector, AI is seldom mentioned in health services. Mentions of AI are the lowest in earnings calls in the communication sector (Figure 9).

Figure 9. Mentions of AI in earnings calls by sector (2018-2019)



Source: Stanford AI Index 2021, <https://aiindex.stanford.edu/report/>.

Annex C. Expert meeting on classifying AI systems, Paris, 27 February 2020

Plenary session: The classification of AI systems

77. ONE AI experts were invited on 27 February to discuss approaches for the classification of AI systems. The meeting's sessions on classifying AI brought together experts and statisticians from the OECD and partner organisations to take stock of existing approaches to classify and quantify AI used by major AI players and statisticians from the OECD, the EC and beyond.

78. Adam Murray, US Delegate to CDEP and Chair of ONE AI moderated the discussion. The Secretariat (Karine Perset and Luis Aranda) introduced the session's goals. Jonathan Frankle, PhD candidate, and Taylor Reynolds, Technology Policy Director, MIT Internet Policy Research Initiative (IPRI) from MIT provided an introduction to AI as well as examples that illustrate the complexity of defining AI. Marko Grobelnik, AI Researcher & Digital Champion at the AI Lab of Slovenia's Jozef Stefan Institute introduced AIGO's work on defining an 'AI system'. Saurabh Mishra, Researcher and Manager of the AI Index Program, Stanford Institute for Human-Centered Artificial Intelligence (HAI), provided insight on lessons learned from the AI index project. Finally, Fernando Galindo-Rueda, Senior Economist, OECD Science and Technology Policy division explained how the OECD has defined and classified other technologies for measurement purposes.

79. Experts also discussed various methodologies and their rationale, commonalities and differences; built on previous OECD work on AI systems and the AI system lifecycle; and took stock of both technical classifications and classifications based on AI systems' application areas.

80. The medium-term (late 2020) aim of this ONE AI work stream is to try to reach a broad agreement on an approach to quantify AI consistently across domains.

Break-out 1: Technical taxonomies

81. Michael Schoenstein, Head of Strategic Foresight & Analysis, Germany moderated the first break-out session on technical taxonomies. Kuansan Wang, Managing Director, Microsoft Research Outreach Academic Services introduced Microsoft Academic Graph (MAG) and the benefits and rationale of using a clustering algorithm to classify automatically types of AI. Giuditta de Prato, Team Leader, EC JRC presented JRC's approach to defining AI for bibliometrics. Dewey Murdick, Director of Data Science at Georgetown University's Center for Security and Emerging Technology discussed AI definitions as used in different measurement contexts.

Break-out 2: Functionalities in AI applications and use cases

82. The second break-out session on functionalities in AI applications and use cases was moderated by Sarah Box, Senior Counsellor, OECD Directorate for Science, Technology and Innovation. Kathleen Walch, managing partner and principal analyst at Cognilytica presented Cognilytica's '7 patterns of AI' of applied AI, standalone AI, or combined AI scenarios. Anand Rao, Partner, Global AI Lead, PwC presented the different time horizon for realising different use cases. Dewey Murdick, Director of Data Science at Georgetown University's Center for Security and Emerging Technology discussed AI definitions as used in different measurement contexts. Pierre Montagnier and Irene Ek policy analysts at OECD Directorate for Science, Technology and Innovation provided an overview of ongoing work taking stock of national AI surveys of AI diffusion among firms.

Annex D. WG CAI Membership

The up-to-date list of WG CAI members and member biographies are available on the OECD.AI Policy Observatory [[link](#)].

Name	Title	Organisation	Group / Delegation
Dewey Murdick (Co-chair)	Director of Data Science	Center for Security and Emerging Technology (CSET), School of Foreign Service, Georgetown University	Civil Society and Academia
Marko Grobelnik (Co-chair)	AI Researcher & Digital Champion	AI Lab of Slovenia's Jozef Stefan Institute	Technical
Jack Clark (Co-chair)	Co-chair	AI Index, Stanford University	Technical
Jefferson de Oliveira Silva	Web Technologies Study Center at NIC.br and Professor of Pontifical Catholic University	NIC.br	Brazil
Sally Radwan	Minister Advisor for Artificial Intelligence	Ministry of Communications & Information Technology (Egypt)	Egypt
Renaud Vedel	Coordonnateur de la stratégie nationale en IA	Ministère de l'intérieur	France
Golo Rademacher	Policy Lab Digital, Work & Society	German Federal Ministry of Labour and Social Affairs	Germany
Judith Peterka	Head, AI indicators	Policy Lab Digital, Work & Society	Germany
Michael Schoenstein	Head of Strategic Foresight & Analysis	Policy Lab Digital, Work & Society	Germany
Barry O'Sullivan	Chair of Constraint Programming, the School of Computer Science & IT	University College Cork	Ireland
David Filip	Research Fellow, ADAPT Centre	Dublin City University (DCU)	Ireland
Yoichi Iida	Chair of the CDEP and Going Digital II Steering Group	Ministry of Internal Affairs and Communications	Japan
Yuki Hirano	Deputy Director, Multilateral Economic Affairs Office, Global Strategy Bureau	Ministry of Internal Affairs and Communications	Japan
Katrina Kosa-Ammari	Counsellor at Foreign Economic Relations Promotion Division	Ministry of Foreign Affairs	Latvia
Olivia Erdelyi	Lecturer/Director of Ethics and Policy	University of Canterbury, School of Law/Soul Machines	New Zealand
Andrey Ignatyev		Ministry of Economic Development	Russia
Anna Abramova	Head of the Department of Digital Economy and Artificial Intelligence	MGIMO-University	Russia
Dunja Mladenčić	Head of Artificial Intelligence Department	Jožef Stefan Institute (Slovenia)	Slovenia
Irene Ek	PhD and leader of the AI portfolio	Swedish Agency for Growth Policy Analysis	Sweden
Bilge Miraç	Advisor	Presidency of Digital Transformation Office (Turkey)	Turkey
Mehmet Haklıdır	Chief Researcher, Scientific and Technological Research Council.	Turkey Informatics and Information Security Research Center	Turkey
Osman Musa Aydın	Advisor to the Deputy Minister, Defense Industry Expert.	Ministry of Industry and Technology	Turkey
Fatma Bujasaim	Head of International Cooperation	Artificial Intelligence Office	United Arab Emirates
Lord Tim Clement-Jones		House of Lords	United Kingdom
Lynne Parker	Deputy United States Chief Technology Officer	The White House	United States
Elham Tabassi	Chief of Staff, Information Technology Laboratory	National Institute of Standards and Technology	United States
Farahnaaz H Khakoo	Assistant Director	US Government Accountability Office	United States

Taka Ariga	Chief Data Scientist Director, Innovation Lab	US Government Accountability Office	United States
Nicholas Reese	Policy expert	Department of Homeland Security	United States
Raj Madhavan	Policy Fellow and Program Analyst	Department of State	United States
Eric Badique	Adviser for Artificial Intelligence	European Commission	European Commission
Emilia Gómez	Lead Scientist, Human behaviour and machine intelligence	European Commission DG Joint Research Centre (JRC)	European Commission
Giuditta de Prato	Researcher	European Commission DG Joint Research Centre (JRC)	European Commission
Prateek Sibal	AI Policy Researcher, Knowledge Societies Division, Communication and Information Sector	UNESCO	IGO
Roberto Sanchez	Advisor - Data Scientist	Inter-American Development Bank	IGO
Alexander Waldmann	Director of Technology & Operations	AppliedAI	Business
Gonzalo López-Barajas Húder	Head of Public Policy and Internet at Telefónica	Telefonica	Business
Igor Perisic	Vice President of Engineering and Chief Data Officer	LinkedIn	Business
Ilya Meyzin	Vice President, Data Science Strategy & Operations	Dun & Bradstreet	Business
Kathleen Walch	Managing partner and principal analyst	Cognilytica	Business
Kuansan Wang	Managing Director	Microsoft Research Outreach Academic Services	Business
Marco Ditta	Executive Director, ISP Group Data Officer	Intesa Sanpaolo	Business
Michel Morvan	Co-Founder and Executive Chairman	Cosmo Tech	Business
Nicole Primmer	Senior Policy Director	BIAC	Business
Nozha Boujemaa	Chief Science & Innovation Officer	Median Technologies	Business
Olly Salzmann	Partner Deloitte/Managing Director	Deloitte KI GmbH and KIParkDeloitte GmbH	Business
Abe Hsuan	Independent Expert	Irwin Hsuan	Technical
Clara Neppel	Senior Director	IEEE European Business Operations	Technical
Jonathan Frankle	PhD Candidate	MIT Internet Policy Research Initiative (IPRI)	Technical
Masashi Sugiyama	Director, Center for Advanced Intelligence Project	RIKEN, Japan	Technical
Peter Addo	Head of DataLab and Senior Data Scientist	Agence Française de Développement (AFD)	Technical
Sebastian Hallensleben	Head of Digitalisation and AI	VDE Association for Electrical, Electronic & Information Technologies	Technical
Taylor Reynolds	Technology Policy Director	MIT Internet Policy Research Initiative (IPRI)	Technical
Catherine Aiken	Researcher	Center for Security and Emerging Technology (CSET), Georgetown University	Civil Society and Academia
Daniel Schwabe	Professor at the Department of Informatics	Catholic University in Rio de Janeiro (PUC-Rio)	Civil Society and Academia
Eva Thelisson	Researcher	AI Transparency Institute	Civil Society and Academia
Guillaume Chevillon	Professor - Co Director ESSEC	ESSEC Business School, Paris	Civil Society and Academia
Jim Kurose	Advisor at the Sorbonne Center for AI	Sorbonne University	Civil Society and Academia
Saurabh Mishra	Researcher and Manager of the AI Index Program	Stanford Institute for Human-Centered Artificial Intelligence (HAI)	Civil Society and Academia
Suso Baleato	Secretary	CSISAC	Civil Society and Academia
Theodoros Evgeniou	Professor, Decision Sciences and Technology Management	INSEAD	Civil Society and Academia
Tim Rudner	PhD Candidate	University of Oxford	Civil Society and Academia

Vincent C. Müller	Professor for Philosophy of Technology	Technical University of Eindhoven	Civil Society and Academia
Yolanda Gil	Director of Knowledge Technologies	University of Southern California	Civil Society and Academia

Secretariat and contact information

The AI Secretariat team within the OECD Digital Economy Policy division supporting this working group are: Karine Perset (Administrator of the OECD’s AI Policy Observatory), Luis Aranda (Policy Analyst, OECD AI Policy Observatory) and Louise Hatem (Junior Policy Analyst, OECD AI Policy Observatory).

Other Secretariat participating in the working group are: Nobuhisa Nishigata (STI/DEP), Marguerita Lane, ELS/SAE (related ELS project under the OECD.AI dedicated Programme on AI in Work, Innovation, Productivity and Skills (AI-WIPS) supported by the German Federal Ministry of Labour and Social Affairs (BMAS)), Fernando Galindo-Rueda, STI/STP (related work on measurement of AI public R&D), Mariagrazia Squicciarini, STI/PIE (related project under the AI-WIPS project), and Pierre Montagnier, STI/DEP (related project under the AI-WIPS project on AI in national surveys of ICT use).

Annex E. Meetings of the OECD Network of Experts Working Group on the Classification of AI Systems

- Plenary and break-out sessions on classifying AI systems at the 27 February 2020 meeting of ONE AI at the OECD in Paris (see Annex C)
- Meeting of 28 May 2020 to discuss the activities and scope of WG CAI [[summary](#)]
- Meeting of 15 June 2020 to finish scoping the work of WG CAI [[scoping document](#)], review [select examples](#) of how AI systems could be classified and discuss the [survey template](#) [[summary](#)].
- Meeting of 2 July 2020 to discuss [survey results](#) and framework’s draft outline
- Meeting of 16 July 2020 to review [analysis of model cards/fact sheets](#) and discuss early draft outlines for the “AI model” and “context” dimensions.
- Meeting of 3 September 2020 to discuss early draft outlines for the “data and input” and “task and output” dimensions
- Meeting of 25 September 2020 to discuss the full outline for the four dimensions and apply the framework to an existing AI system
- Meeting of 22 October 2020 to discuss the interim report to CDEP
- Meeting of 2 December 2020 to discuss preliminary CDEP feedback and next steps
- Meeting of 16 December 2020 to informally brainstorm and test the framework on selected real-life examples of AI systems.
- Meeting of 19 January 2021 to discuss feedback received on the framework from delegates, parliamentarians and working group members’ respective expert communities.
- Meeting of 15 February 2021 to review and discuss revisions brought to the model dimension of the framework and share members’ experience in using the framework.
- Meeting of 2 March 2021 to present preliminary work on the sectoral impacts of AI and related policy considerations.

Notes

¹ A smaller steering group composed of the co-chairs, the Secretariat and consultants meets regularly between WG sessions.

² ‘Data and input’ *includes* ‘data collection’ to simplify the framework. In previous OECD work, c.f. Box 1, the ‘Data collection’ and ‘Data / input’ itself were separate objects.

³ ‘Task and output’ *includes* ‘Action’ to simplify the framework. In previous OECD work, c.f. Box 1, the AI system ‘Output’ and its ‘Action’ were separate objects.

⁴ 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 (OECD, 2019^[11]).

⁵ 1) Context: system users; 2) Data and input: actors collecting and processing data and input; 3) AI model: system developers; and, 4) Tasks and output: system operators –

⁶ The OECD AI Principles say “or decisions”, which the expert group decided should be excluded to clarify that an AI system does not make an actual decision, which is the remit of human creators and outside the scope of the AI system.

⁷ The characterisation of an AI system has been adapted by replacing the term ‘interpret’ with ‘use’ to avoid confusion with the term ‘model interpretability.’

⁸ <https://gdpr-text.com/read/article-22/#links>

⁹ 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 (OECD, 2019^[11]).

¹⁰ A number of applications combine symbolic and statistical approaches. For example, natural language processing (NLP) algorithms often combine statistical approaches (that build on large amounts of data) and symbolic approaches (that consider issues such as grammar rules). Autonomous driving systems for example use both machine-based inputs (historical driving data) and human-based inputs (a set of driving rules). Combining models built on both data and human expertise is viewed as promising to help address the limitations of both approaches.

¹¹ Abrams M, “The Origins of Personal Data and its Implications for Governance” [\[Link\]](#)

¹² <https://www.stateof.ai/> based on research by OpenAI.

¹³ Both the data itself and its format can be proprietary, *i.e.* the data format could be known only by the owner.

¹⁴ <https://www.ftc.gov/news-events/media-resources/protecting-consumer-privacy-security/ftc-policy-work>

¹⁵ See for example <https://www.sciencedirect.com/science/article/pii/S2052297520301104>.

¹⁶ Federated learning is a subset of distributed machine learning. While distributed machine learning runs algorithms on edge devices to split the training workload across different machines, federated learning trains the algorithm on the edge, summarises the changes and only sends this focused update back to the main model.

¹⁷ The uncertainty can be structural (e.g. is a linear regression or a neural network more appropriate, and if the latter, how many layers should it have etc.), parametric (which values of a model’s parameters make the best prediction) or noise-related (e.g., pixel noise or blur in images).

¹⁸ While machine learning techniques are usually used in building or adjusting a model, they can also be used to interpret a model’s results.

¹⁹ This section leverages and builds on research undertaken by Cognilytica on the “7 patterns of AI”.

²⁰ <https://www.3ds.com/sustainability/sustainability-insights/designing-disruption/executive-summary>

²¹ This draws on the ongoing work of the OECD Employment, Labour and Social Affairs Directorate (ELS) and the Global Partnership on AI (GPAI) Working Group on the Future of Work.

²² <http://www.oecd.org/education/trustworthy-artificial-intelligence-in-education.pdf>

References

- AI Ethics Impact Group (2020), *From Principles to Practice: An interdisciplinary framework to operationalise AI ethics*, <https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf>. [8]
- Alekseeva, L. (2019), “The demand for AI skills in the labour market.”, *Center for Economic Policy Research (CEPR)*, <https://repec.cepr.org/repec/cpr/ceprdp/DP14320.pdf>. [20]
- CISA (2019), *National Critical Functions Set*, Cybersecurity & Infrastructure Security Agency, <https://www.cisa.gov/national-critical-functions-set>. [5]
- CoE (2020), *The feasibility study on AI legal framework adopted by the CAHAI*, <https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da>. [10]
- CoE (1998), *European Convention on Human Rights*, https://www.echr.coe.int/documents/convention_eng.pdf. [11]
- EC (2020), *WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust*, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. [3]
- Endsley (1987), *The application of human factors to the development of expert systems for advanced cockpits*. [19]
- Hao, K. (2019), *Training a single AI model can emit as much carbon as five cars in their lifetimes*, MIT Technology Review, <https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/>. [18]
- Mankins, J. (1995), *Technology readiness levels*, White Paper, <https://www.sciencedirect.com/science/article/pii/S0736585320301842#bb0035>. [7]
- Martinez Plumed, F. (2020), *AI Watch: Assessing Technology Readiness Levels for Artificial Intelligence*, Publications Office of the European Union, Luxembourg, <http://dx.doi.org/doi:10.2760/15025>. [6]
- OECD (2020), *How’s Life? 2020: Measuring Well-being*, OECD Publishing, Paris, <https://doi.org/10.1787/9870c393-en>. [13]

- OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, <https://dx.doi.org/10.1787/eedfee77-en>. [2]
- OECD (2019), *Enhancing Access to and Sharing of Data*. [16]
- OECD (2019), *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies*, OECD Publishing, Paris, <https://doi.org/10.1787/276aaca8-en>. [15]
- OECD (2019), *Recommendation of the Council on Artificial Intelligence*, <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>. [21]
- OECD (2019), *Recommendation of the Council on Digital Security of Critical Activities*, OECD, <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0456>. [4]
- OECD (2019), “Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)”, *OECD Digital Economy Papers*,, https://www.oecd-ilibrary.org/science-and-technology/scoping-the-oecd-ai-principles_d62f618a-en. [14]
- OECD (2019), “Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)”, *OECD Digital Economy Papers*, <https://doi.org/10.1787/d62f618a-en>. [11]
- OECD (2013), *The OECD Privacy Framework*, https://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf. [22]
- OECD (2011), *The Role of Internet Intermediaries in Advancing Public Policy Objectives*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264115644-4-en>. [9]
- OECD (Forthcoming), *Draft Report on the Implementation of the OECD Recommendation concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*. [23]
- OHCHR (2011), *Guiding Principles on Business and Human Rights*, United Nations Human Rights Office of the High Commissioner, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf. [12]
- Russell, S. (2019), *Human-compatible*, Penguin Books, http://ISBN_9780525558637. [17]