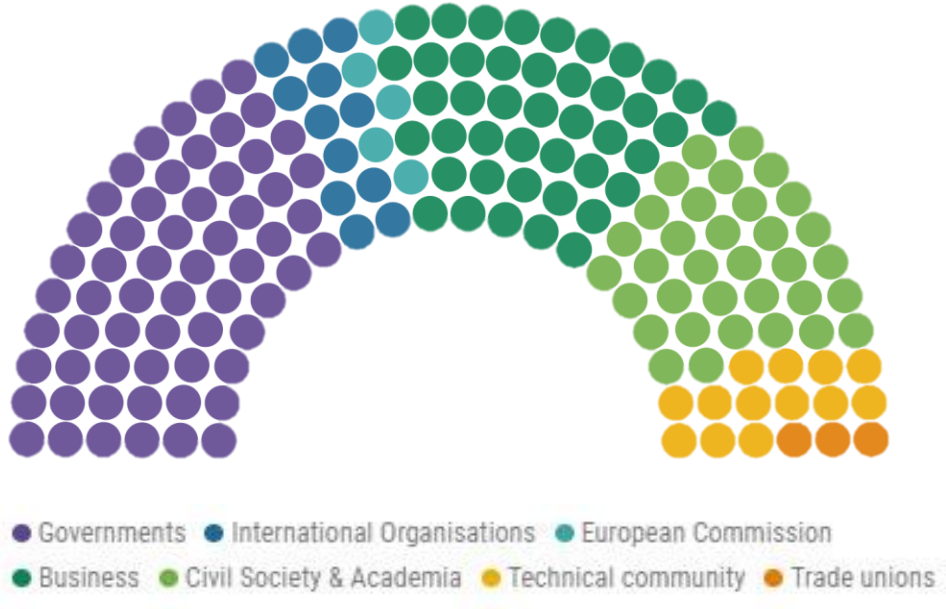
The OECD Framework for Classifying AI systems

Introducing the OECD Network of Experts on AI (ONE AI)

- Provides **AI-specific expertise** and advice on implementing the **OECD AI Principles**
- Launched in **February 2020**
- **200+ AI experts** from national governments, IGOs and the EC, business, civil society, academia, trade unions
- Facilitates **collaboration** between the OECD and other international initiatives on AI



Three working groups and one task force to implement the **OECD AI Principles**

OECD.AI Policy Observatory

Working Group on Classifying AI systems

Values-based principles
Socio-economic & environmental impacts
Human-centred values and fairness
Transparency, explainability
Robustness, security, safety
Accountability

National Policies
Investing in research
Compute, data, technologies
Enabling policy environment
Jobs, skills, transitions
International cooperation

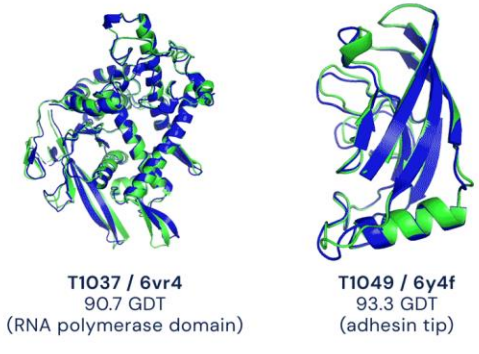
Task Force on AI Compute

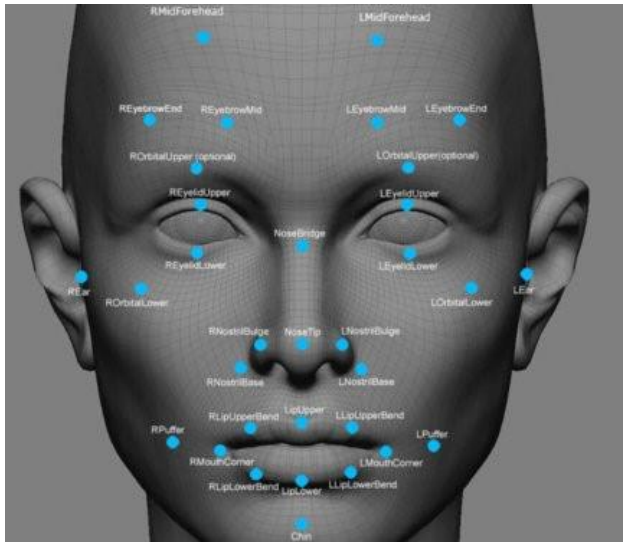
WG on Tools for Trustworthy AI

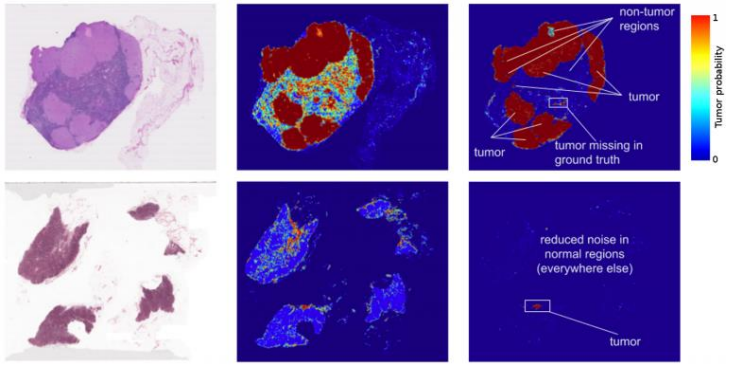
WG on National AI policies

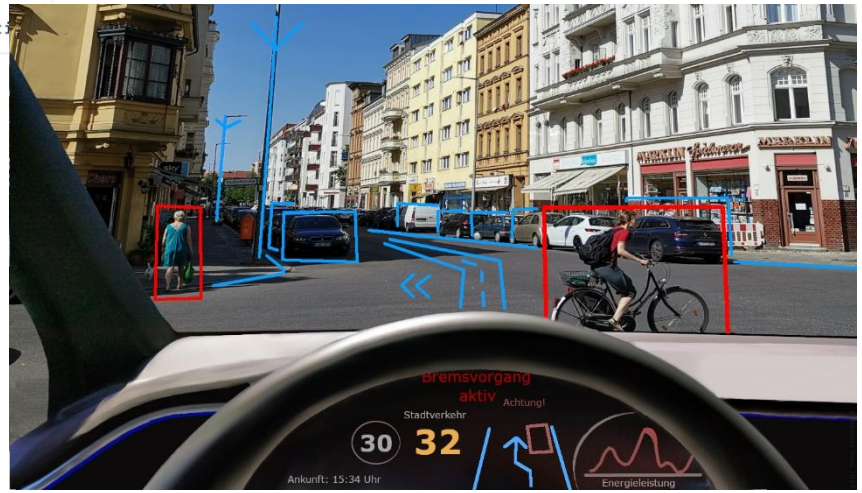
Why classify AI systems?

A variety of systems and policy implications









OECD AI System Definition (OECD, 2019)

"An **AI system**, is a machine-based system that is capable of **influencing the environment** by producing an **output** (recommendations, predictions or decisions) for a given set of **objectives**.

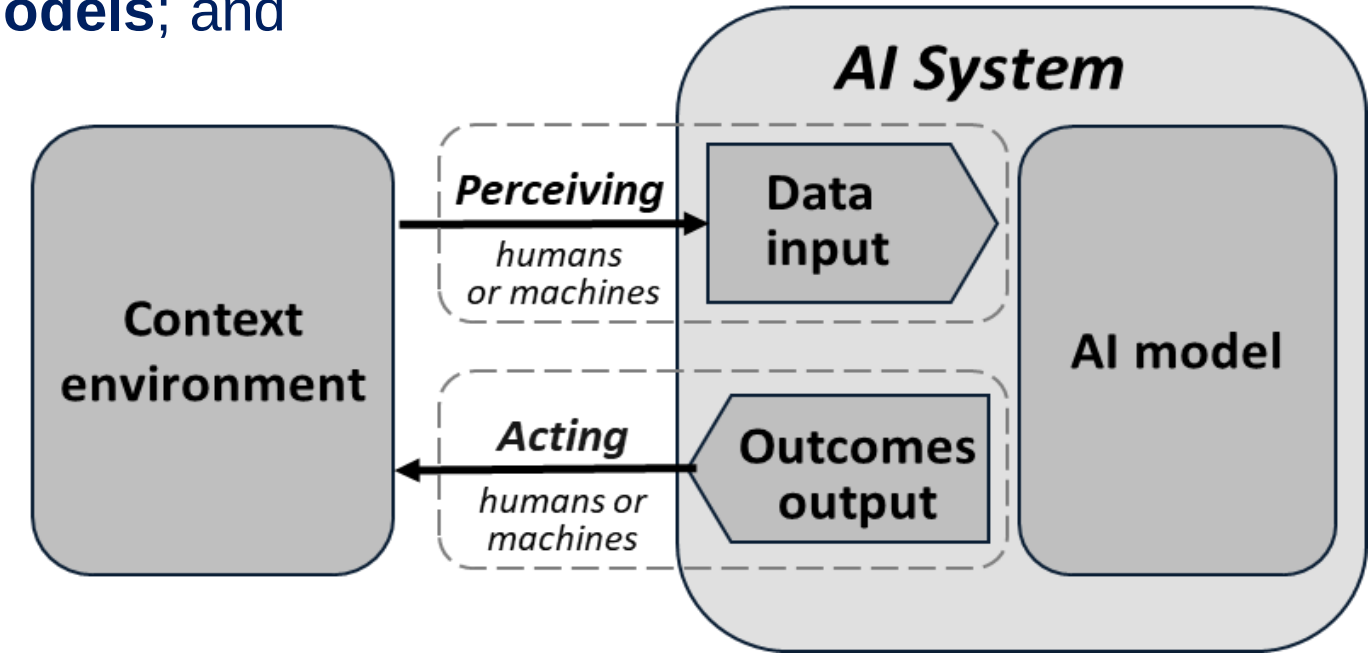
It uses machine and/or human-based inputs/data to:

i) **perceive** environments;

ii) abstract these perceptions **into models**; and

iii) **use** the models to formulate options for **outcomes**.

AI systems are designed to operate with **varying levels of autonomy.**"



What is the OECD Framework for the Classifying AI systems?

- A **structure to differentiate AI systems** & map **implications for key policy areas** framed by the OECD AI Principles
- A **user friendly and simplified tool** : 4 dimensions and 20 core criteria
- Linking **technical aspects to policy considerations**

The Framework has 4 dimensions & 20 core criteria

OECD.AI Policy Observatory

1. CONTEXT

- Industrial sector
- Business function
- Critical function
- Scale and maturity
- Users
- Impacted stakeholders, optionality, business model
- Human rights impact
- Well-being impact

Key actors include: system operators and end users

2. DATA AND INPUT

- Provenance, collection and dynamic nature
- Structure and format (structured etc.)
- Rights and 'identifiability' (personal, proprietary etc.)
- Appropriateness and quality

Key actors include: data collectors and processors

3. AI MODEL

- Model characteristics
- Model building (symbolic, machine learning, hybrid)
- Model inferencing / use

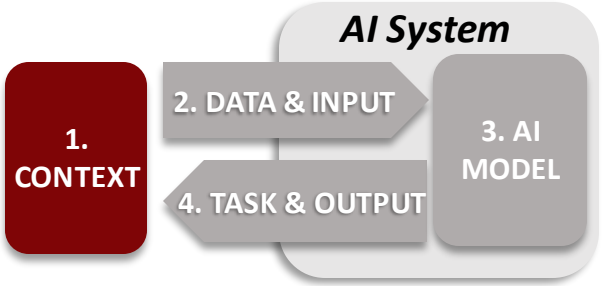
Key actors include: developers and modellers

4. TASK AND OUTPUT

- Task of the system (recognition; personalisation etc.)
- Action of the system (autonomy level)
- Combining tasks and action
- Core application areas (computer vision etc.

Key actors include: system integrators

1. Context of the AI system



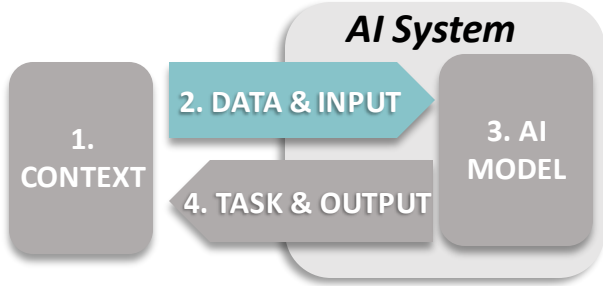
CONTEXT

Which industry?	Associated industry structure & regulation, *e.g.,* healthcare, transportation, public admin.
For what business function?	*e.g.,* production; human resources management; sales;
How widely deployed?	Pilot stage, narrow, broad or wide deployment
What type of users?	AI experts or amateurs
Impacting whom?	*e.g.,* consumers; employees; businesses; government; groups.
With what degree of choice?	Whether users can or cannot opt out and / or challenge system output
What impact on human rights?	System's potential on fundamental human rights (e.g; liberty; freedom of expression)
Performs a critical function?	System performs critical function or is deployed in critical activity sector
How mature?	Defined using NASA TRLs
What business model?	For profit – advertising, subscription, other --; non-profit
What impact on well-being?	System's potential on well-being (e.g. on health, on job quality)

Key actors include:
system operators and end users

Core criteria
grey : proposed as optional criteria that may be difficult to answer

2. Data & input



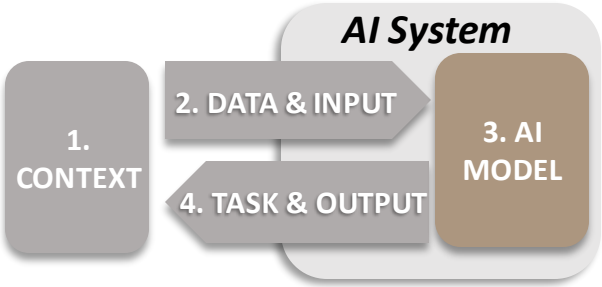
DATA AND INPUT

Where does the data come from?	Expert input, provided data, observed data, synthetic data, derived data
How is it collected?	By humans, by automated sensing devices, or both
How dynamic is it?	Static data, dynamic updated from time to time, dynamic real time
Is the data structured or not?	Unstructured, semi-structured, structured, complex structured
Are rights attached to the data / input?	Proprietary, public or personal data
What is dataset's scale?	Small, medium, large, very large
What is the data's format?	Standardised or non-standardised
Is the data identifiable?	Identified, pseudonymised, anonymised
Is the data "appropriate"?	Criteria to ensure that data are fit for the purpose
What is the data's quality?	Representative sample and sufficient sample size, completeness, data noise

Core criteria

grey : proposed as optional criteria that may be difficult to answer

3. AI Model

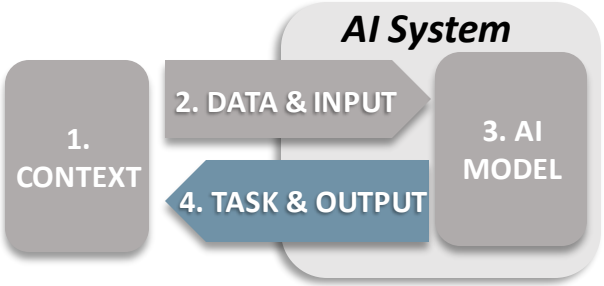


AI MODEL

What is the type of model used?	Symbolic, statistical or hybrid
How is the model trained / built?	Human written-rules, supervised, semi-supervised, unsupervised learning;
Does the model evolve & how?	No evolution, evolution through passive interaction, evolution through active interaction in machine learning models
Is it discriminative? generative?	Discriminative (i.e. predicting data labels by learning to distinguish between dataset classes) or generative model (i.e. explaining how the data was generated)
Does the model learn and how?	Centralised or federated learning
How is the model used?	Deterministic (one outcome), probabilistic model (of distribution of outcomes) or both

Core criteria
grey : proposed as optional criteria that may be difficult to answer

4. Task and output



TASK AND OUTPUT

What is the task performed by the system?	Recognition, event detection, forecasting, personalisation, interaction support, goal-driven optimisation, reasoning w. knowledge structures
How autonomous are the system's actions and what role do humans play?	No action autonomy, low, medium, high autonomy; Human support, human in-the-loop, on-loop, out-of-the loop
What is the labour displacement potential of the system?	Potential for displacement through task automation
Does the system belong to a core application area?	Human language technologies, computer vision, robotics
Does the system combine different tasks & actions ?	Autonomous system, control system, content generation

Core criteria
grey : proposed as optional criteria that may be difficult to answer

Work in progress: Mapping systems' classification to risk

- Identified **characteristics that signify a system is not low risk,** including **potential risks for human rights,** building on Council of Europe's work
- Associated other characteristics with **positive, negative or neutral impact on risk** to obtain a preliminary cumulative effect

Next steps:

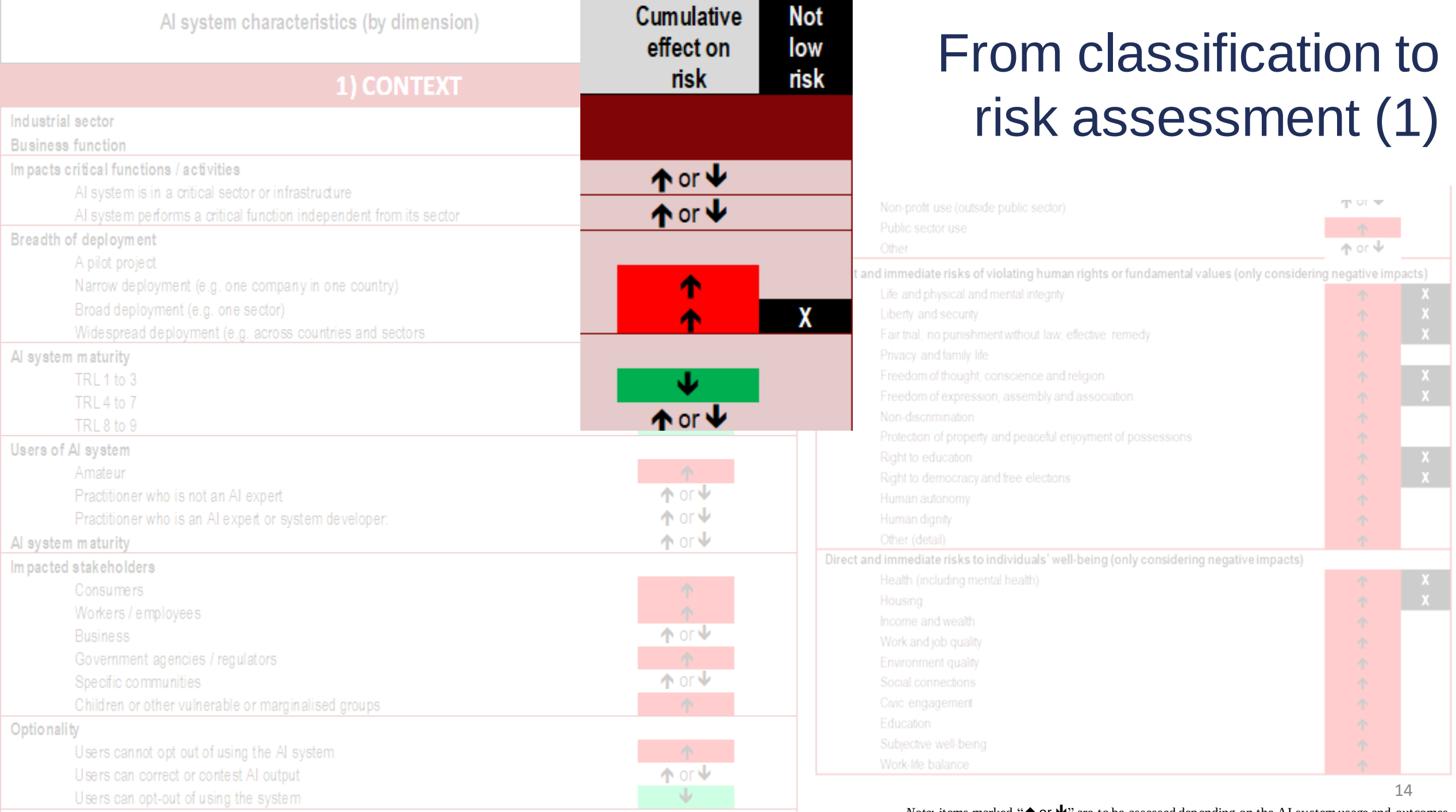
- Refine methodology and define output (*eg*. composite score)
- Test relevance & applicability

From classification to risk assessment (1)

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		
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	↑	X
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	↑	

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

From classification to risk assessment (1)



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

From classification to risk assessment (2, 3, 4)

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

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 ↓

Example 1: AI-powered hiring system

OECD.AI Policy Observatory

1. CONTEXT

Sector: Administrative and support services

Business function: Human resources

Critical function: No

Scale: Narrow

System users: Amateur (job candidate and HR)

Impacted stakeholders: applicants / employees

Human rights impact: YES

Well-being impact: YES

2. DATA AND INPUT

Data provenance: Provided by candidate; Observed by algorithms; Derived

Collection: Collected by automated tools

Dynamism: Static; dynamic; real-time

Structure : Unstructured data

Rights & Identifiability: Proprietary personal data; identified

Appropriateness & quality: Unknown

3. AI MODEL

Model type: Hybrid

Model building / training: Semi-supervised

Discriminative model

Model inferencing / use: Probabilistic

4. TASK AND OUTPUT

Task: Personalization, interaction support, recognition

Autonomy level: Medium – provides recommendation for human decision execution

Composite system: Yes

Belongs to core application area: TBD

Example 2: Credit scoring system

OECD.AI Policy Observatory

1. CONTEXT

Sector: Financial and insurance activities

Business function: Sales & customer service

Critical function: YES

Scale: Broad

System users: Non-AI expert (bank employee)

Impacted stakeholders: consumers

Optionality: None

Human rights impact: YES
Well-being impact: YES

2. DATA AND INPUT

Provenance: expert rules; observed by algorithms (eg. payments history); provided by consumer (eg, name); derived (eg. credit score)

Collection: By humans & by sensing devices

Dynamism: Static (eg, gender), dynamic and real-time (eg. payments)

Structure & Format: structured & standardised

Rights: Personal, proprietary, identified data

Appropriate for purpose **; Quality unknown**

3. AI MODEL

Model type: Hybrid (statistical & symbolic)

Model building / training: Acquisition from data, Augmented by human-encoded knowledge

Learning mode: Central learning

Model inferencing / use: Deterministic

4. TASK AND OUTPUT

Task: Forecasting & reasoning with knowledge structures

Autonomy level: Low

Composite system: YES (combines tasks and actions)

Belongs to core application area: TBD

Example 3: Manufacturing plant management system

OECD.AI Policy Observatory

1. CONTEXT

Sector: Construction

Business function: Sales, logistics, HR, monitoring …

Critical function: No

Scale: Narrow

Maturity: TRL 9

System users: Amateurs, non-expert and experts

Impacted stakeholders: Workers, consumers, firms

Optionality: Variable

Human rights impact: YES
Well-being impact: YES

2. DATA AND INPUT

Provenance: all (expert rules; observed by algorithms ; provided by consumer; derived ; synthetic)

Collection: by humans & by sensing devices

Dynamism: Static (human knowledge), dynamic real-time (from production lines)

Structure & Format: all types of structures; standardized and un-standardized format

Rights: Proprietary

Appropriate for purpose **; Representative** and appropriate, noise /missing values/outliers known

3. AI MODEL

Model type: Hybrid (statistical & symbolic)

Model building / training: Acquisition from data, Augmented by human-encoded knowledge

Learning mode: Federated

Data interaction: Evolution during operation through active and passive interaction

Model inferencing / use: both deterministic & probabilistic

4. TASK AND OUTPUT

Task: All (forecasting; recognition; reasoning…)

Autonomy level: Medium

Displacement potential: High

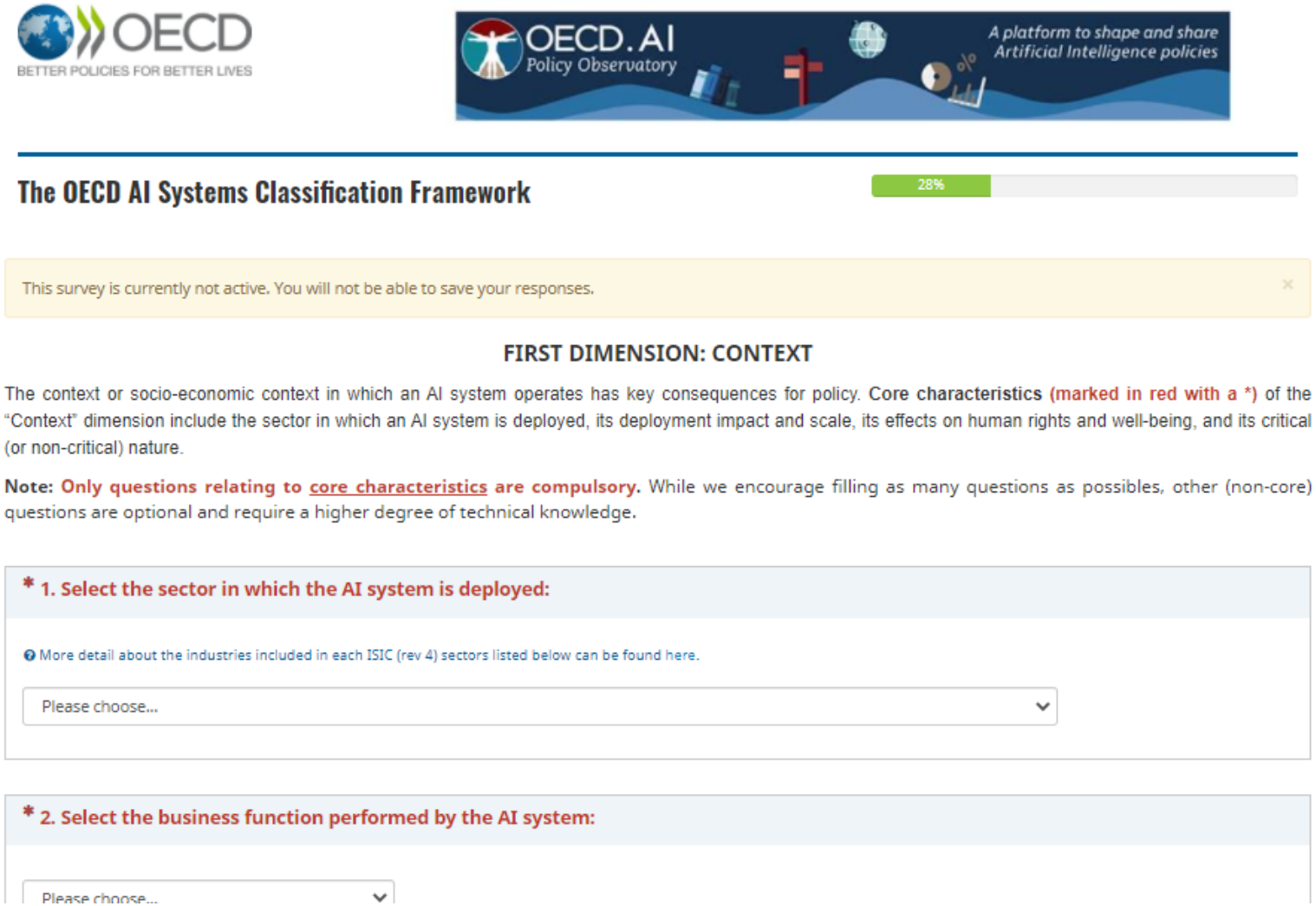
Composite system: Yes

Belongs to core application area: Human language technologies (as well as process planning and optimisation and Internet-of-things)

How to participate in the consultation?

➡ oecd.ai/p/classification

1. **Comment on report :** provide comments and suggestions directly on the written version of the Framework.
2. **Take our online survey to test the Framework** on an AI system of your choice.



Public consultation on the Framework

Key questions

1. Should **core and and non-core criteria** be distinguished? I.e. should there be a core classification framework for less-technical audiences + additional considerations for more technical / informed users?
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 are useful for a **more detailed, technically-oriented** framework?
5. Should there be **industry or application domain specific criteria** and classifications, e.g. depending on context?

Timeline for the consultation

➡ *oecd.ai/p/classification*

1. **Timeline**: 20 May 2021 – 30 June 2021
2. **Comments summarised and published**: summer 2021
3. **Planned launch for the Framework:** fall 2021