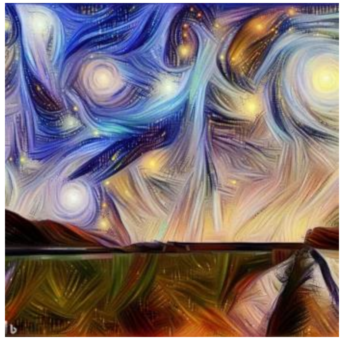
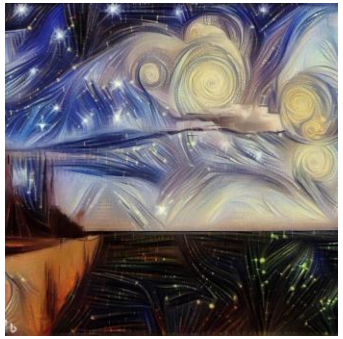



OECD AI Policy Framework & Some Thoughts on AI

Marko Grobelnik

(Marko.Grobelnik@ijs.si)

Artificial Intelligence Department, Jozef Stefan Institute

UNESCO IRCAI

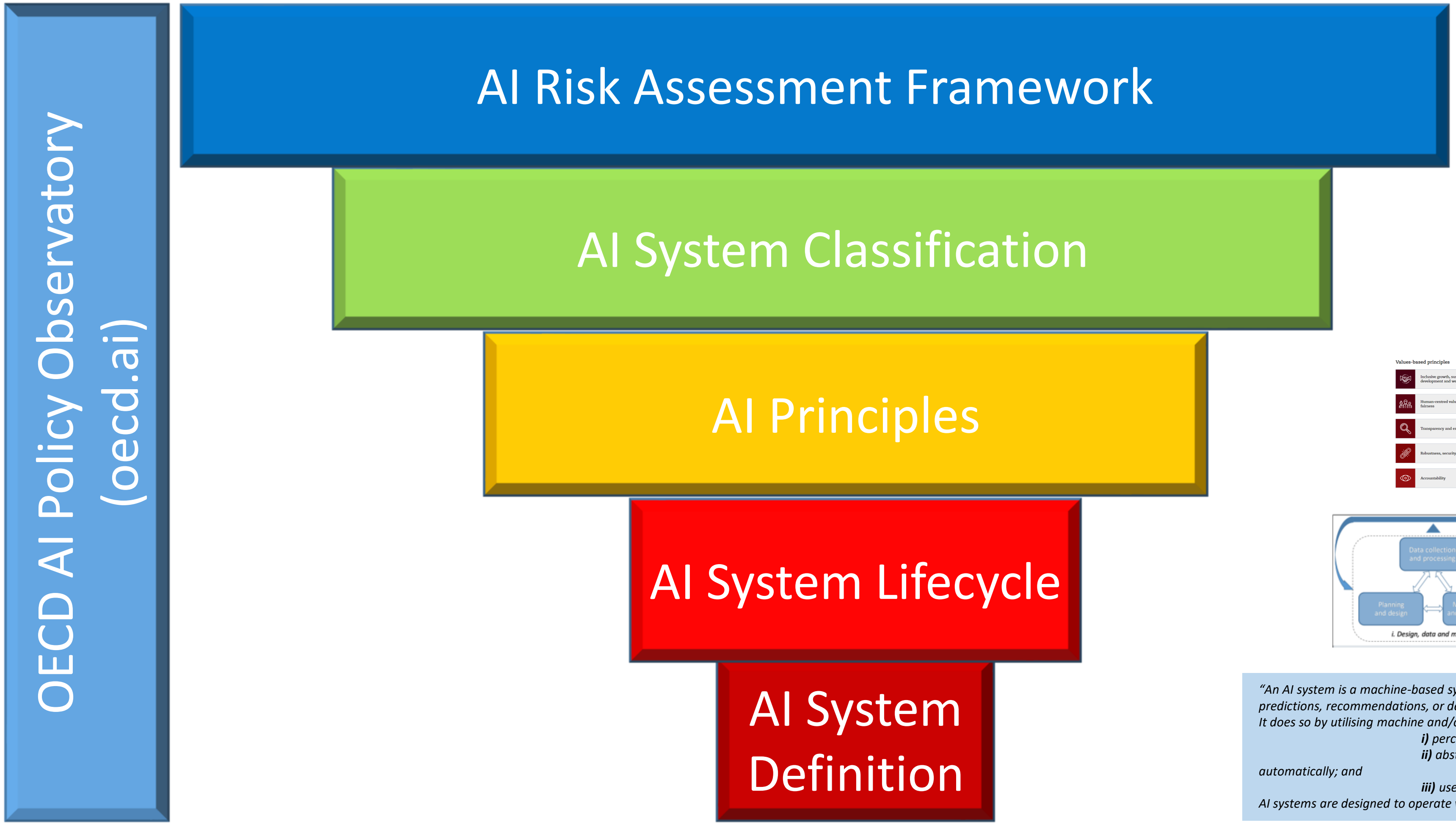
Institut
"Jožef Stefan"
Ljubljana, Slovenija

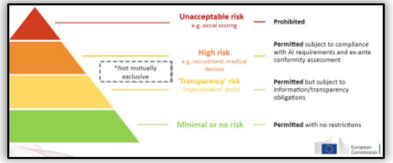
United Nations
Educational, Scientific and
Cultural Organization

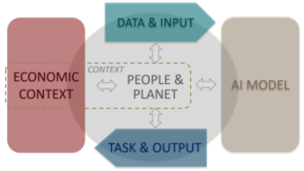
International Research Centre
on Artificial Intelligence
under the auspices of UNESCO

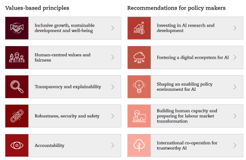
The Context:
OECD AI Policy Framework

OECD AI Policy Framework – the overall schema

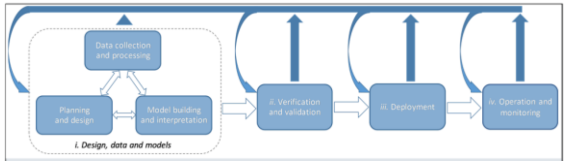



Unacceptable risk
e.g. social scoring
Prohibited
High risk
e.g. recruitment, medical devices
Permitted subject to compliance with AI requirements and ex-ante conformity assessment
*Not mutually exclusive
'Transparency' risk
'Impersonation' (bots)
Permitted but subject to information/transparency obligations
Minimal or no risk
Permitted with no restrictions
European Commission

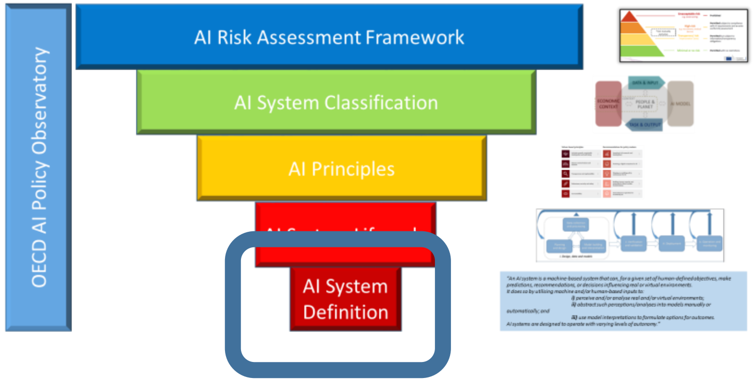




Values-based principles	Recommendations for policy makers
Inclusive growth, sustainable development and well-being	Investing in AI research and development
Human-centred values and fairness	Fostering a digital ecosystem for AI
Transparency and explainability	Shaping an enabling policy environment for AI
Robustness, security and safety	Building human capacity and preparing for labour market transformation
Accountability	International co-operation for trustworthy AI


Data collection and processing
Planning and design
Model building and interpretation
i. Design, data and models
ii. Verification and validation
iii. Deployment
iv. Operation and monitoring

"An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.
It does so by utilising machine and/or human-based inputs to:
***i)** perceive and/or analyse real and/or virtual environments;*
***ii)** abstract such perceptions/analyses into models manually or automatically; and*
***iii)** use model interpretations to formulate options for outcomes.*
AI systems are designed to operate with varying levels of autonomy."



"An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. It does so by utilising machine and/or human-based inputs to: i) perceive and/or analyse real and/or virtual environments; ii) abstract such perceptions/analyses into models manually or automatically; and iii) use model interpretations to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy."

OECD AI System Definition

Adopted in 2019

Name of the game: *Definition of AI*

Informal definition of non-AI

- AI is exactly the **opposite** from what is happening in the video…
- …instead of living beings mimicking machines, AI is intended to make machines imitating living beings.

AI Definitions from the literature

- *“The exciting new effort to make computers think…[as] machines with minds, in the full and literal sense.” (Haugeland 1985)*
- *“[The automation of] activities that we associate with human thinking such as decision-making, problem-solving, learning.” (Bellman 1978)*
- *“The art of creating machines that perform functions that require intelligence when performed by people.” (Kurzweil 1990)*
- *“The study of how to make computers do things which, at the moment, people are better.” (Rich and Knight 1991)*
- *“The study of the computations that make it possible to perceive, reason, and act.” (Winston 1992)*
- *“Making machines intelligent; intelligence is that quality that enables an entity to function appropriately and with foresight in its environment.” (Nils Nilsson)*

OECD AI Definition (OECD 2019)

(adopted also by G20 and EC)

“An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.

It does so by utilising machine and/or human-based inputs to:

***i)** perceive and/or analyse real and/or virtual environments;*

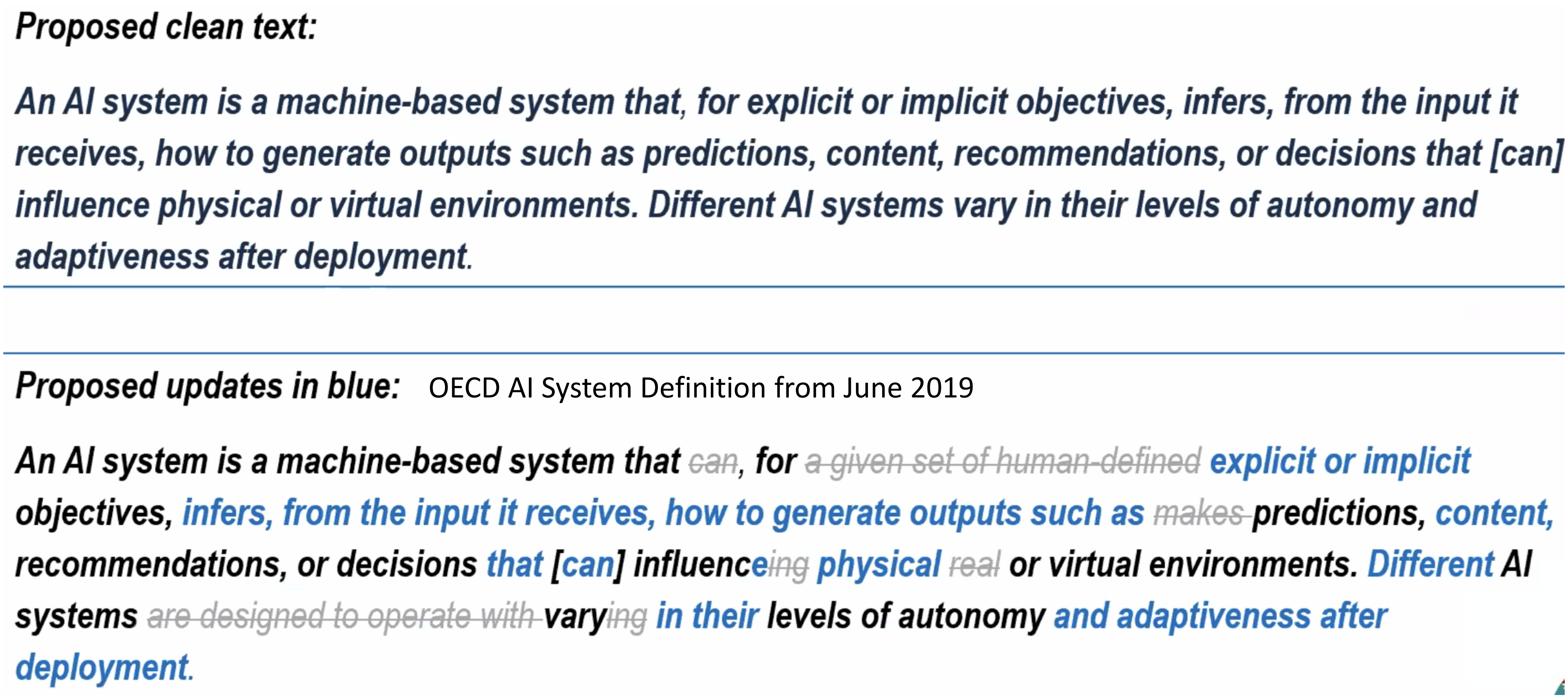
***ii)** abstract such perceptions/analyses into models manually or automatically; and*

***iii)** use model interpretations to formulate options for outcomes.*

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

New OECD AI System definition (Oct 16th 2023)

(adopted by EU AI Act, G7, US NIST, Council of Europe)



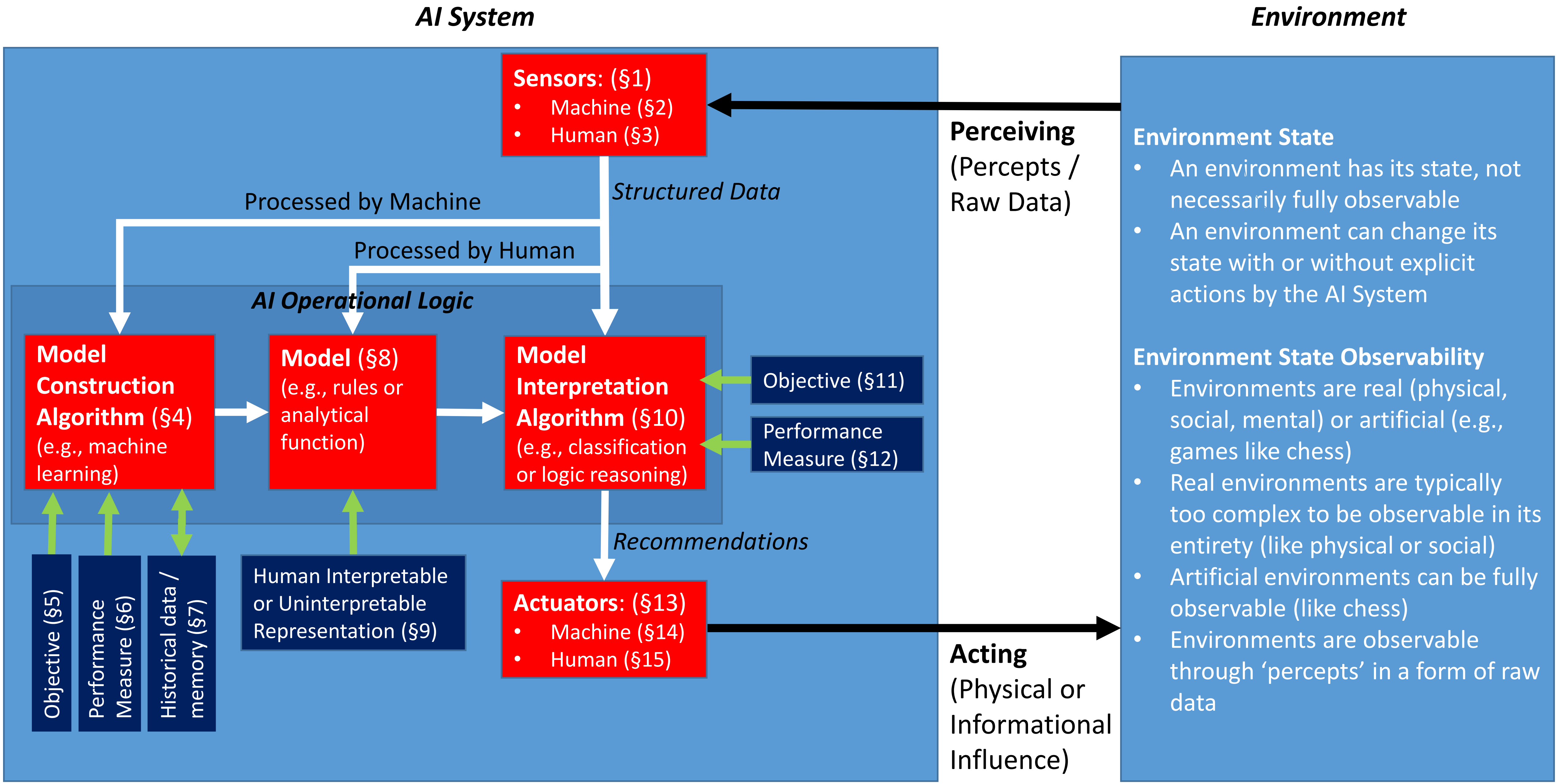
Proposed clean text:

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

Proposed updates in blue: OECD AI System Definition from June 2019

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

Anatomy of the AI System definition (as defined by OECD)



AI System

Environment

Sensors: (§1)
- Machine (§2)
- Human (§3)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

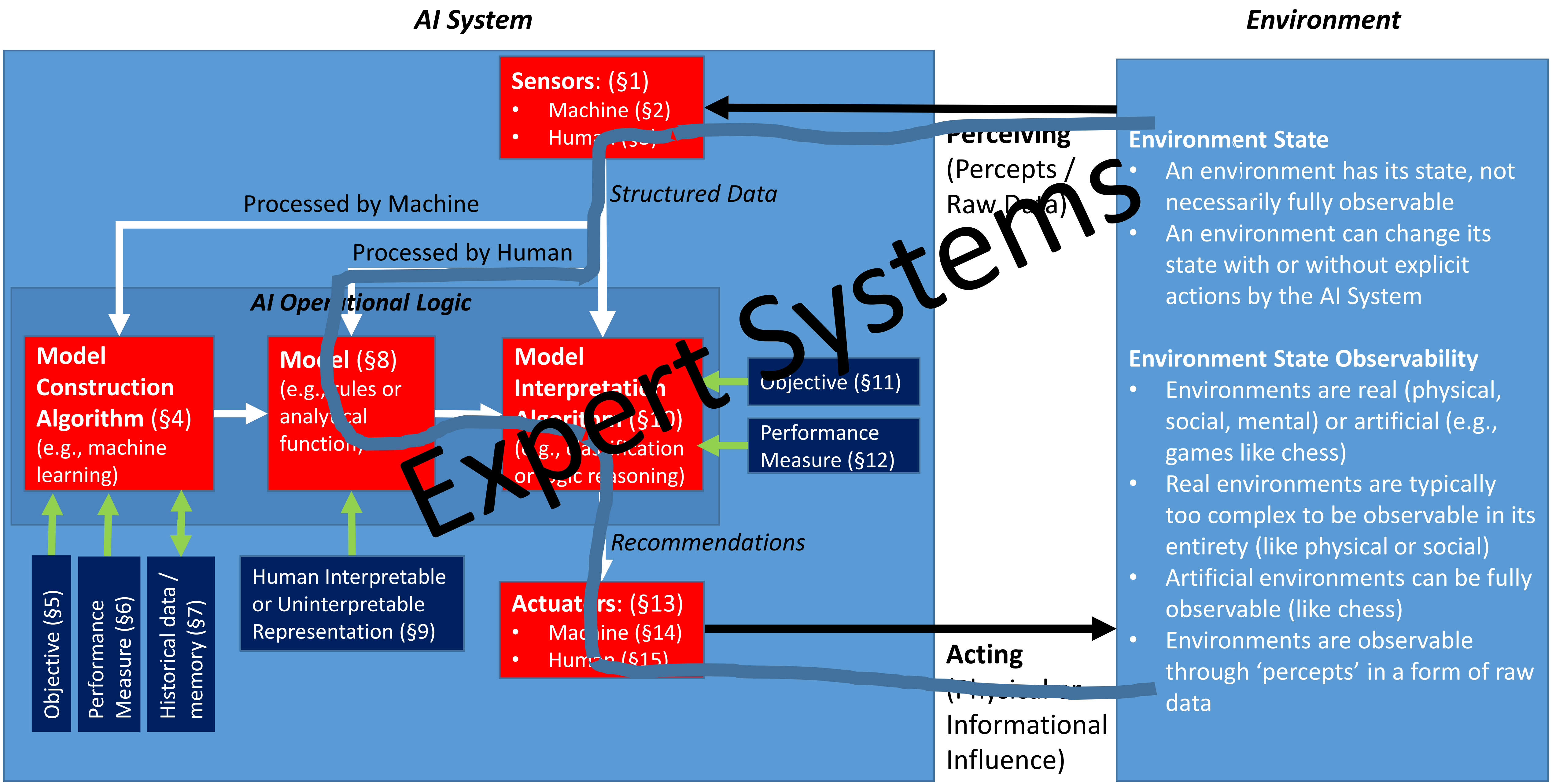
Perceiving (Percepts / Raw Data)

Acting (Physical or Informational Influence)

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through ‘percepts’ in a form of raw data

AI System as defined by OECD



AI System

Environment

Sensors: (§1)
- Machine (§2)
- Human (§3)

Processed by Machine

Processed by Human

Structured Data

Perceiving

(Percepts / Raw Data)

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

Acting

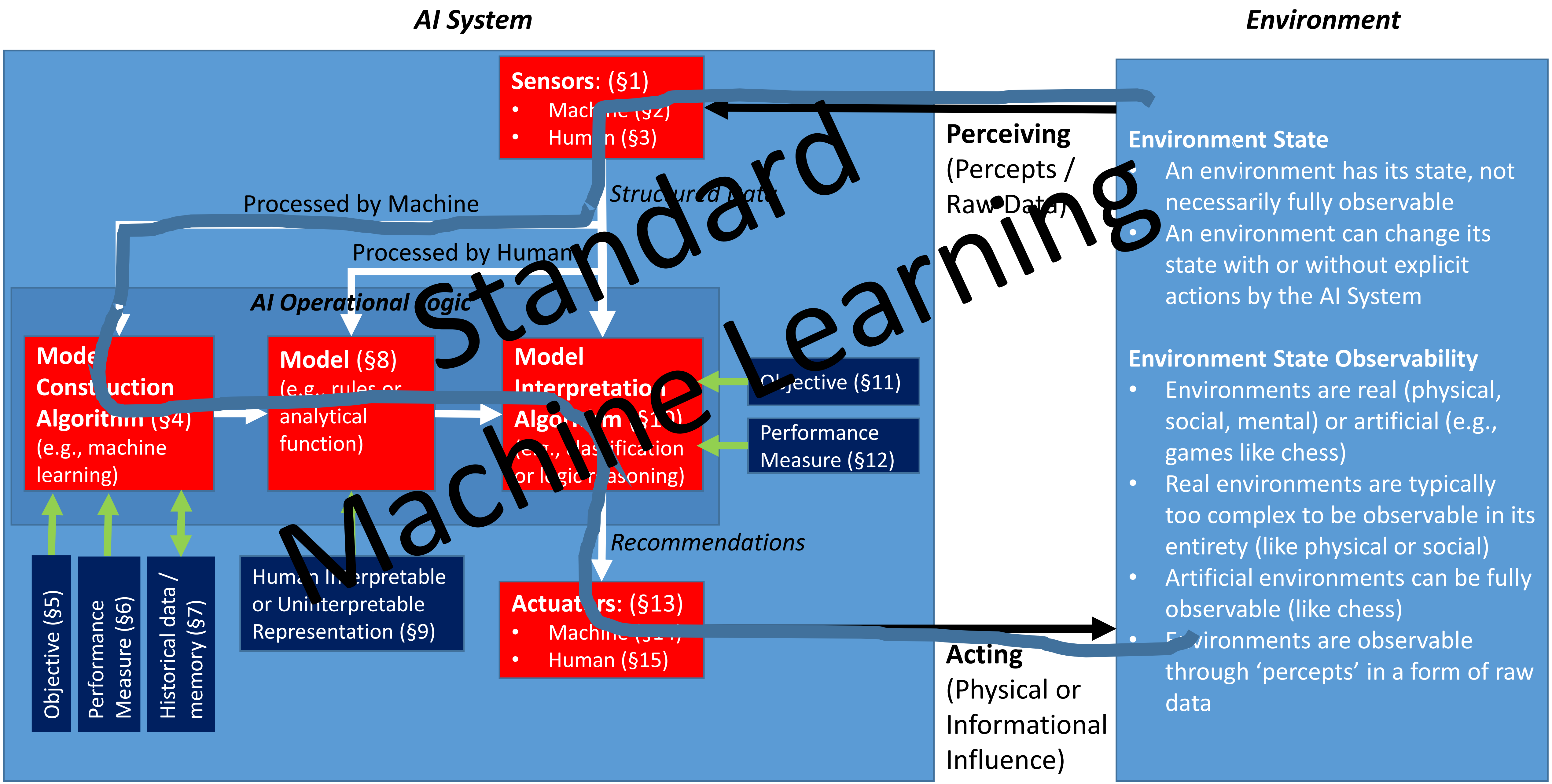
(Physical or Informational Influence)

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through ‘percepts’ in a form of raw data

Expert Systems

AI System as defined by OECD



AI System

Environment

Sensors: (§1)
- Machine (§2)
- Human (§3)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

Perceiving (Percepts / Raw Data)

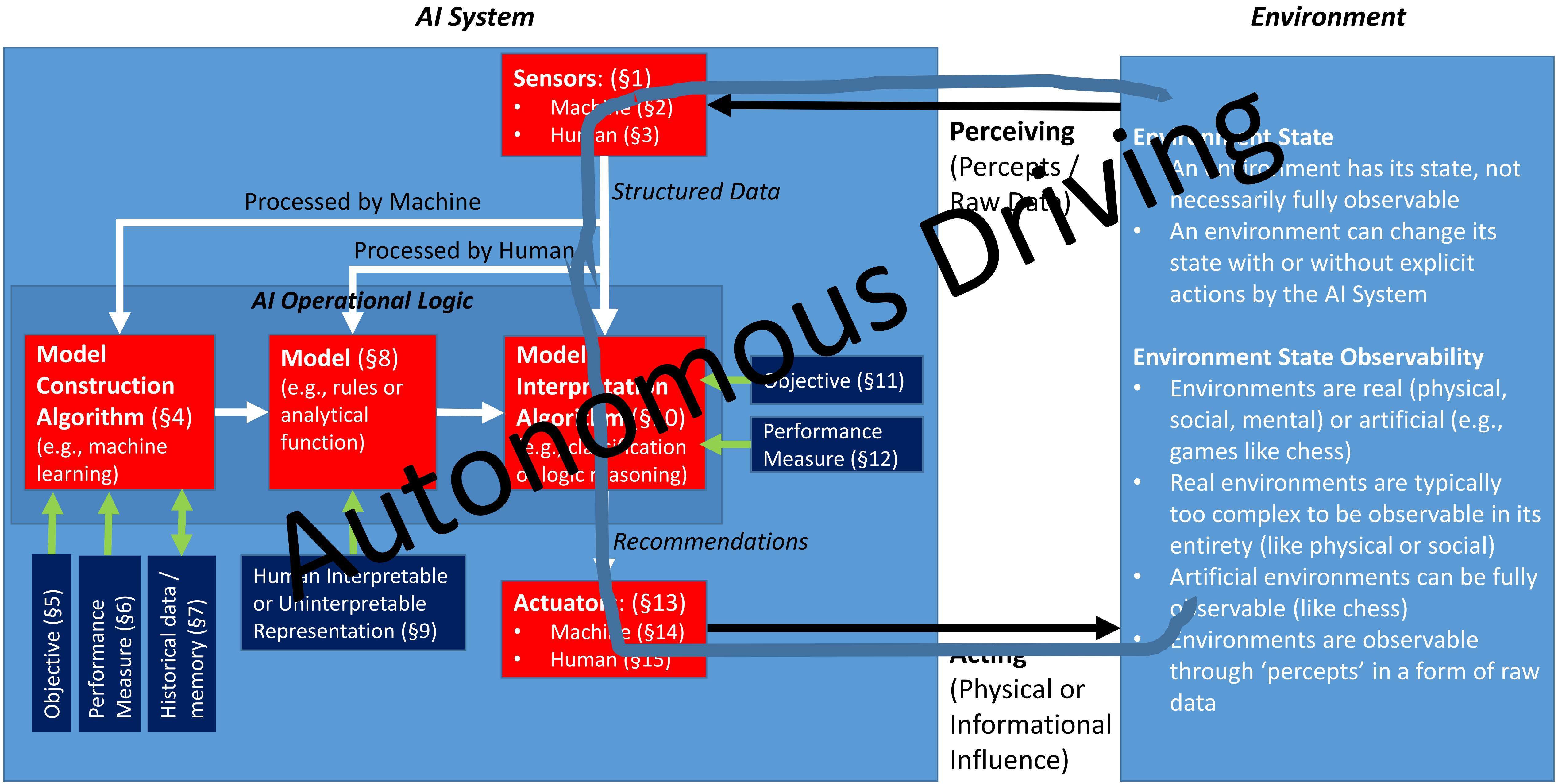
Acting (Physical or Informational Influence)

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through ‘percepts’ in a form of raw data

Standard Machine Learning

AI System as defined by OECD



AI System
Environment
Sensors: (§1)
Machine (§2)
Human (§3)
Structured Data
Processed by Machine
Processed by Human
AI Operational Logic
Model Construction Algorithm (§4) (e.g., machine learning)
Model (§8) (e.g., rules or analytical function)
Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)
Objective (§11)
Performance Measure (§12)
Objective (§5)
Performance Measure (§6)
Historical data / memory (§7)
Human Interpretable or Uninterpretable Representation (§9)
Recommendations
Actuators: (§13)
Machine (§14)
Human (§15)
Perceiving (Percepts / Raw Data)
Acting (Physical or Informational Influence)
Autonomous Driving

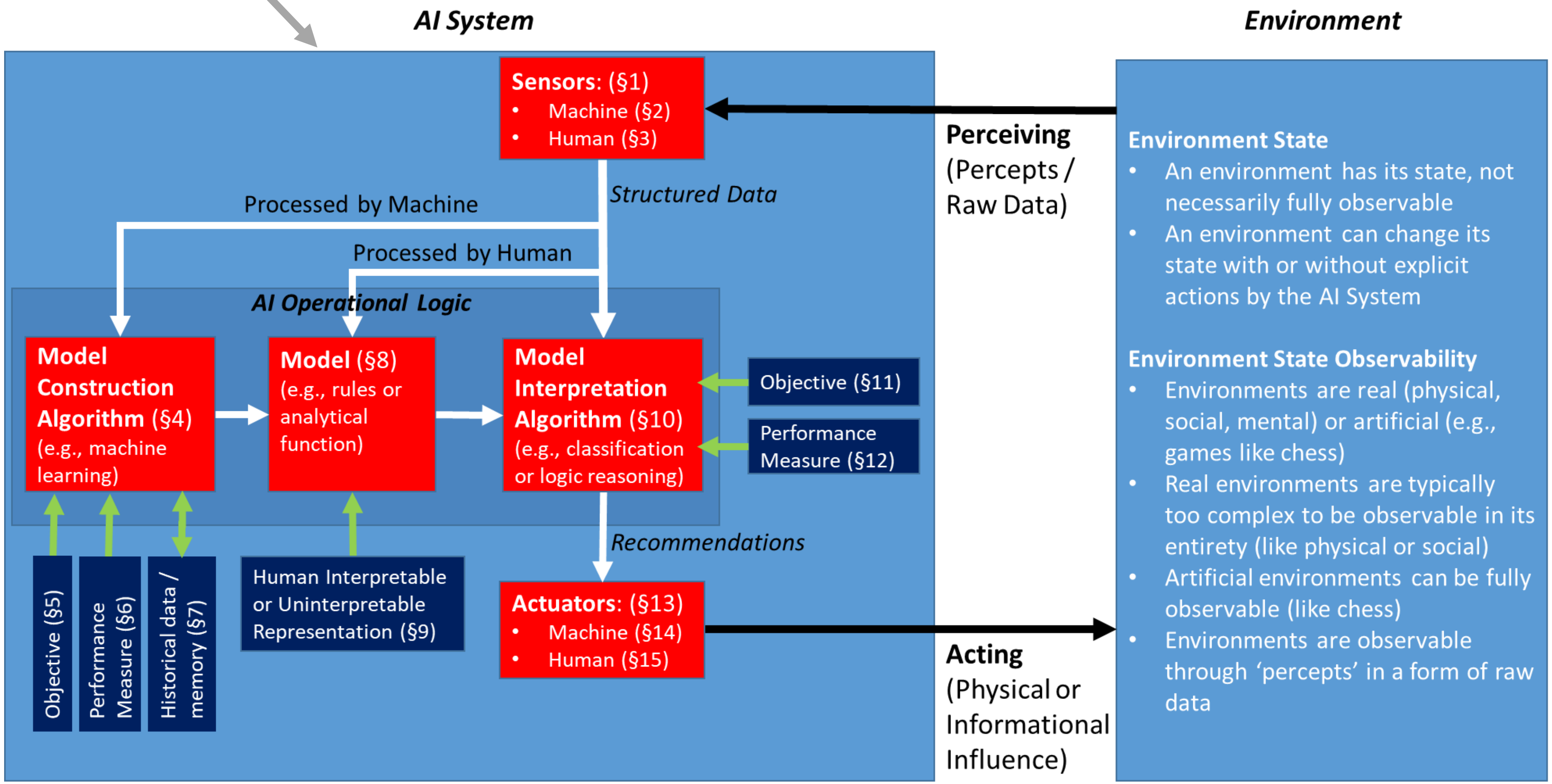
Environment State

- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability

- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through ‘percepts’ in a form of raw data

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



AI System

Environment

Sensors: (§1)
- Machine (§2)
- Human (§3)

Perceiving
(Percepts / Raw Data)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

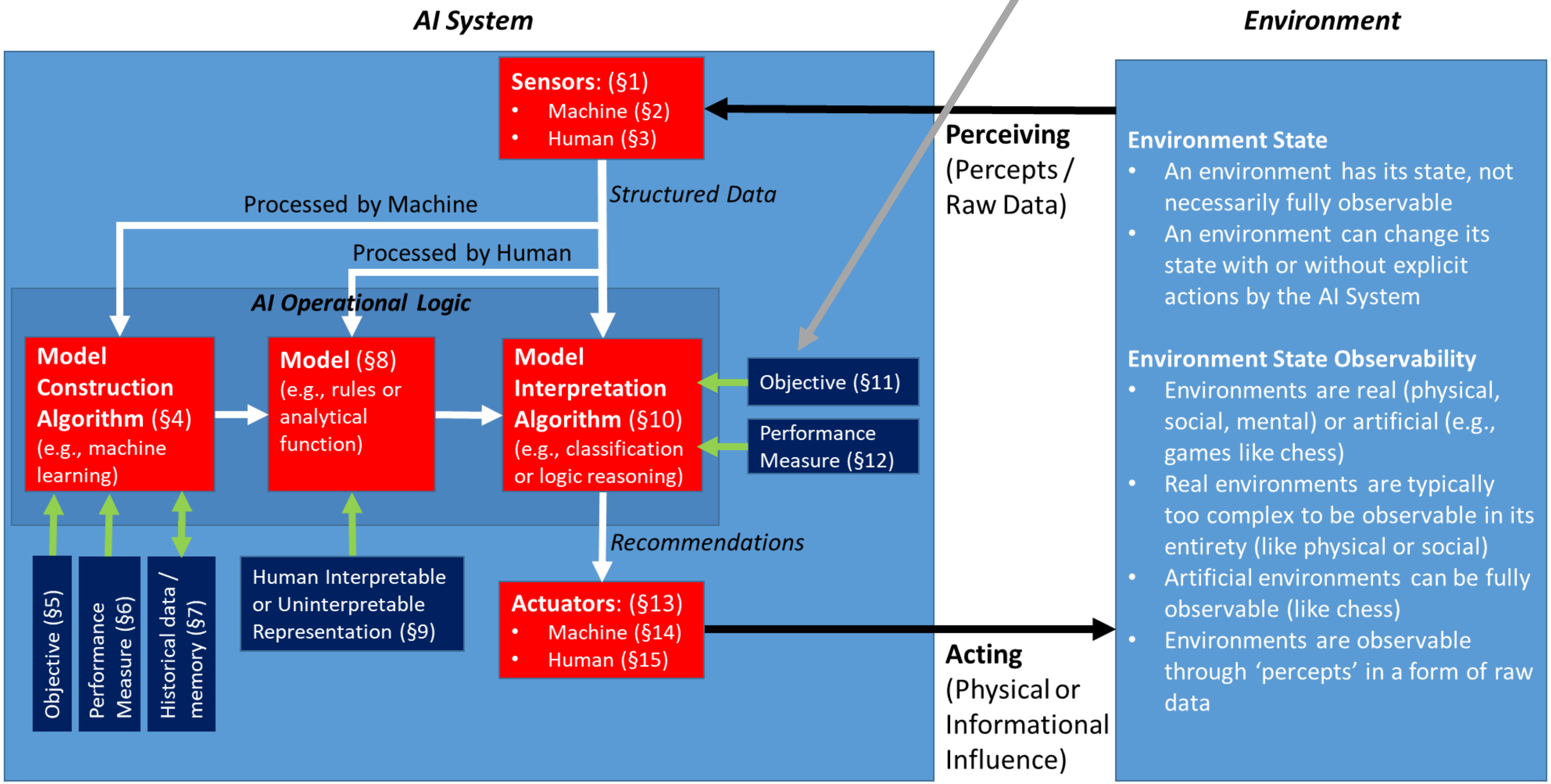
Actuators: (§13)
- Machine (§14)
- Human (§15)

Acting
(Physical or Informational Influence)

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through 'percepts' in a form of raw data

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


AI System
Environment
Sensors: (§1)
• Machine (§2)
• Human (§3)
Structured Data
Processed by Machine
Processed by Human
Perceiving
(Percepts / Raw Data)
AI Operational Logic
Model Construction Algorithm (§4) (e.g., machine learning)
Model (§8) (e.g., rules or analytical function)
Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)
Objective (§11)
Performance Measure (§12)
Objective (§5)
Performance Measure (§6)
Historical data / memory (§7)
Human Interpretable or Uninterpretable Representation (§9)
Recommendations
Actuators: (§13)
• Machine (§14)
• Human (§15)
Acting
(Physical or Informational Influence)

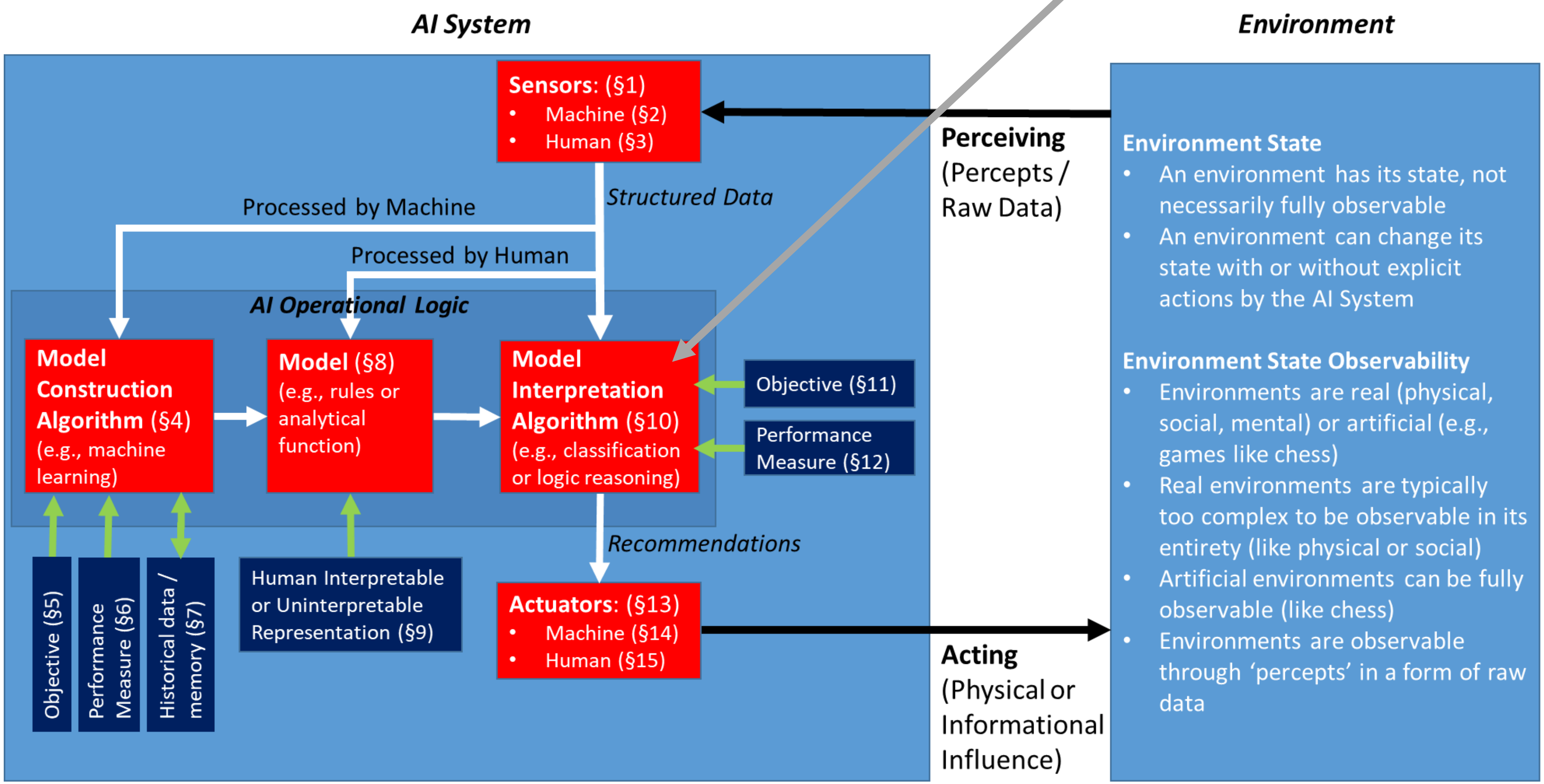
Environment State

- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability

- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through ‘percepts’ in a form of raw data

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



AI System

Sensors: (§1)
- Machine (§2)
- Human (§3)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

Perceiving (Percepts / Raw Data)

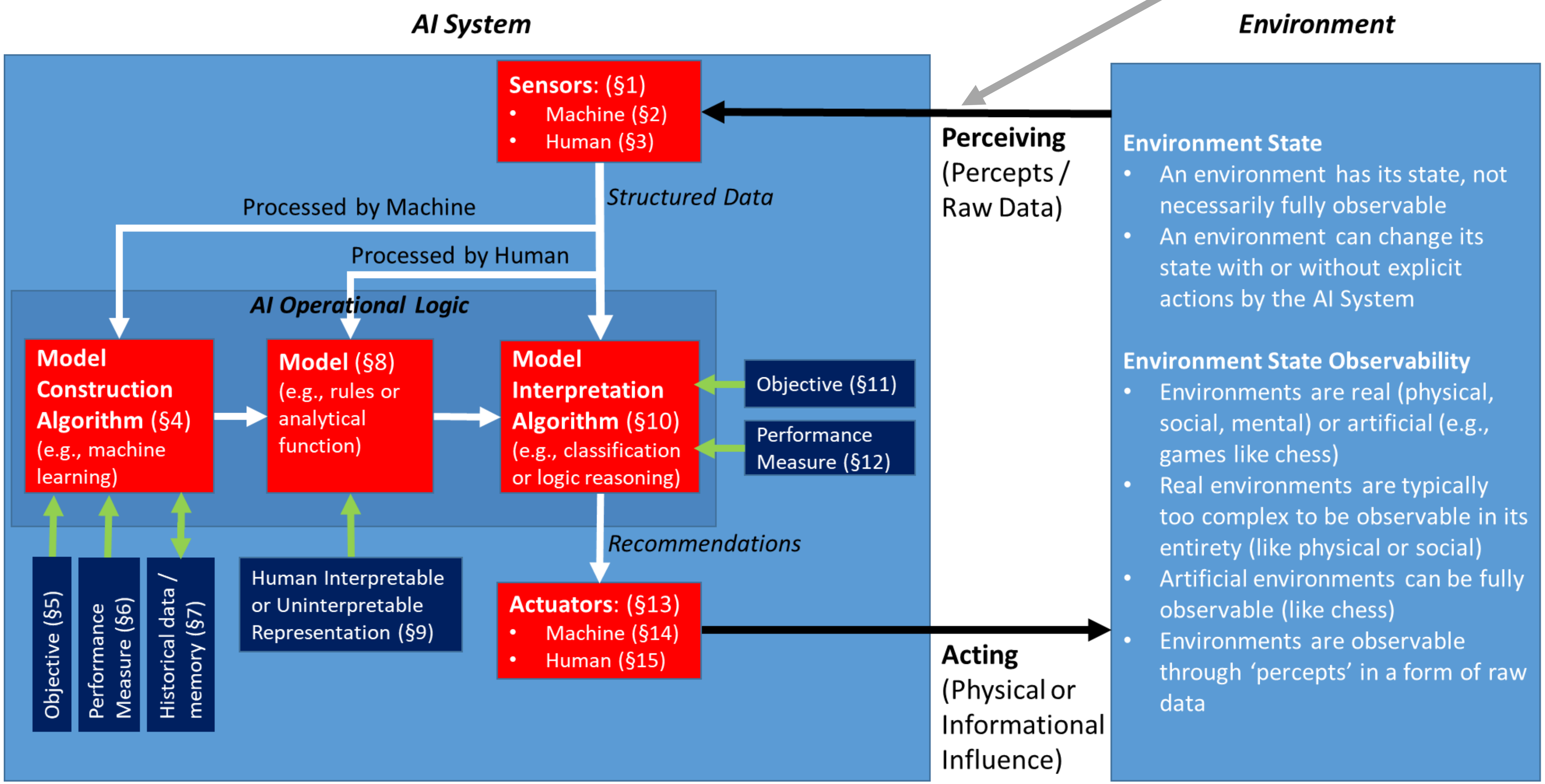
Acting (Physical or Informational Influence)

Environment

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through ‘percepts’ in a form of raw data

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



AI System

Sensors: (§1)
- Machine (§2)
- Human (§3)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

Perceiving (Percepts / Raw Data)

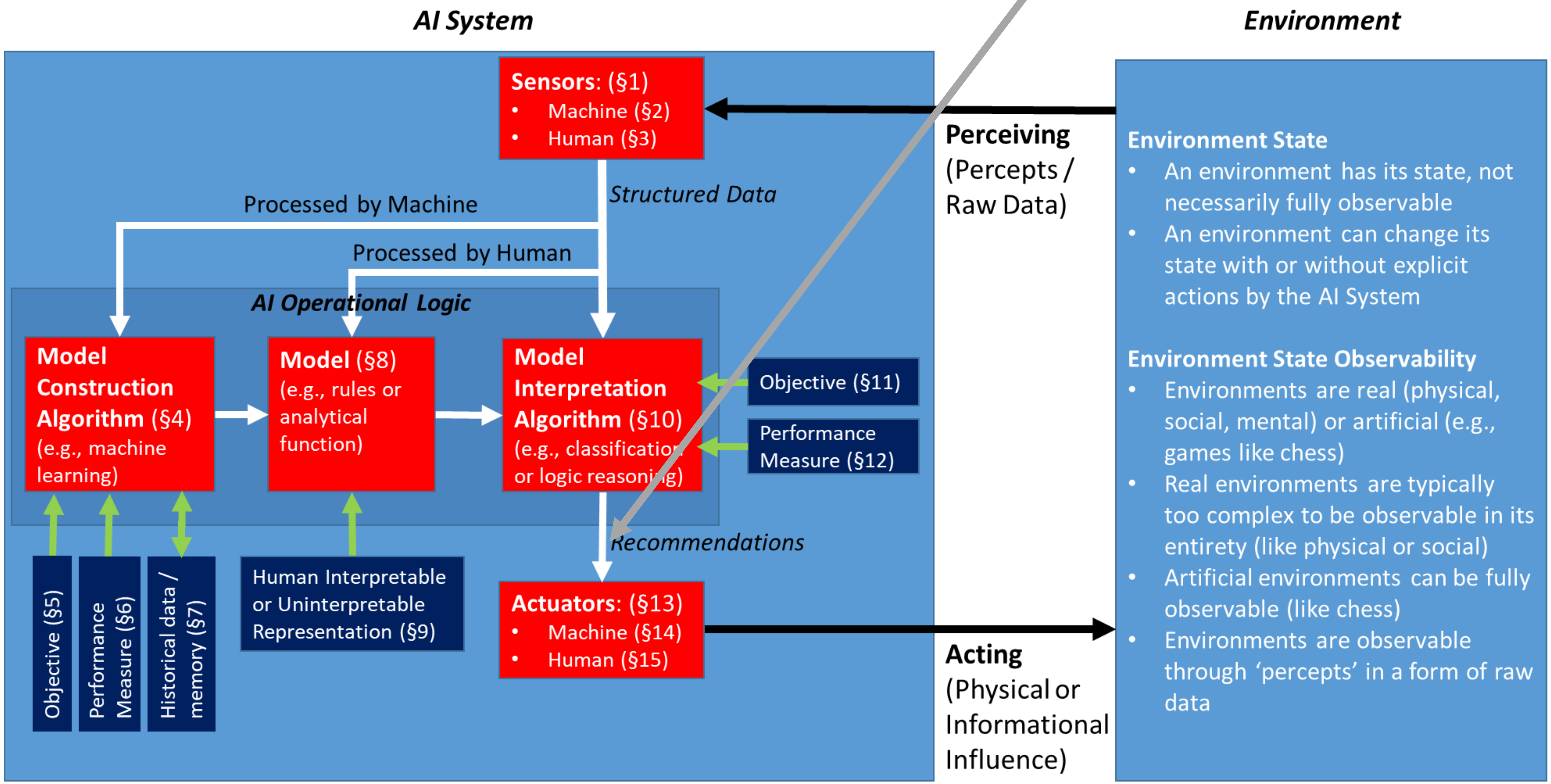
Acting (Physical or Informational Influence)

Environment

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through 'percepts' in a form of raw data

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



AI System

Sensors: (§1)
- Machine (§2)
- Human (§3)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

Perceiving (Percepts / Raw Data)

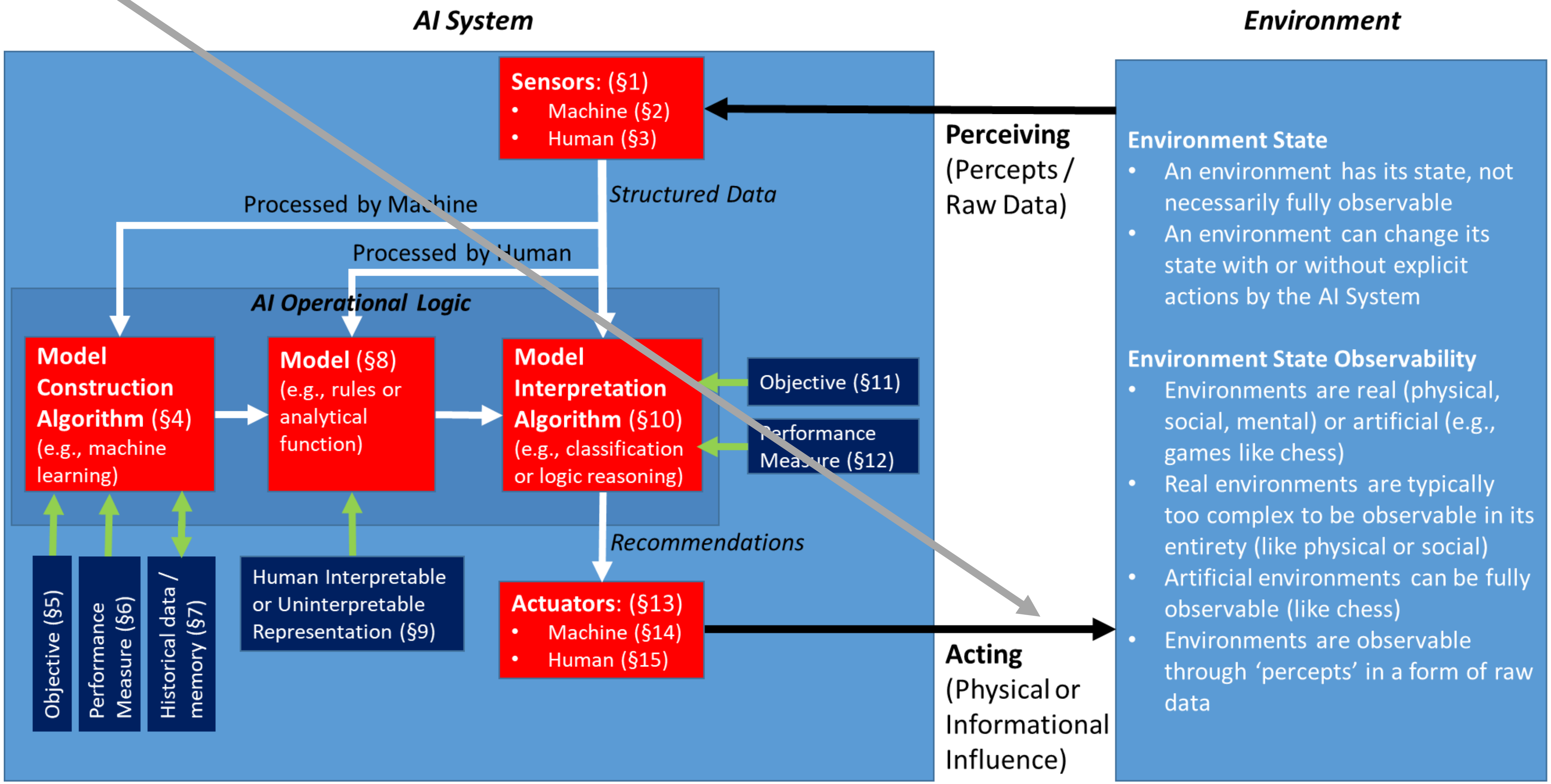
Acting (Physical or Informational Influence)

Environment

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through ‘percepts’ in a form of raw data

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



AI System

Sensors: (§1)
- Machine (§2)
- Human (§3)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

Perceiving (Percepts / Raw Data)

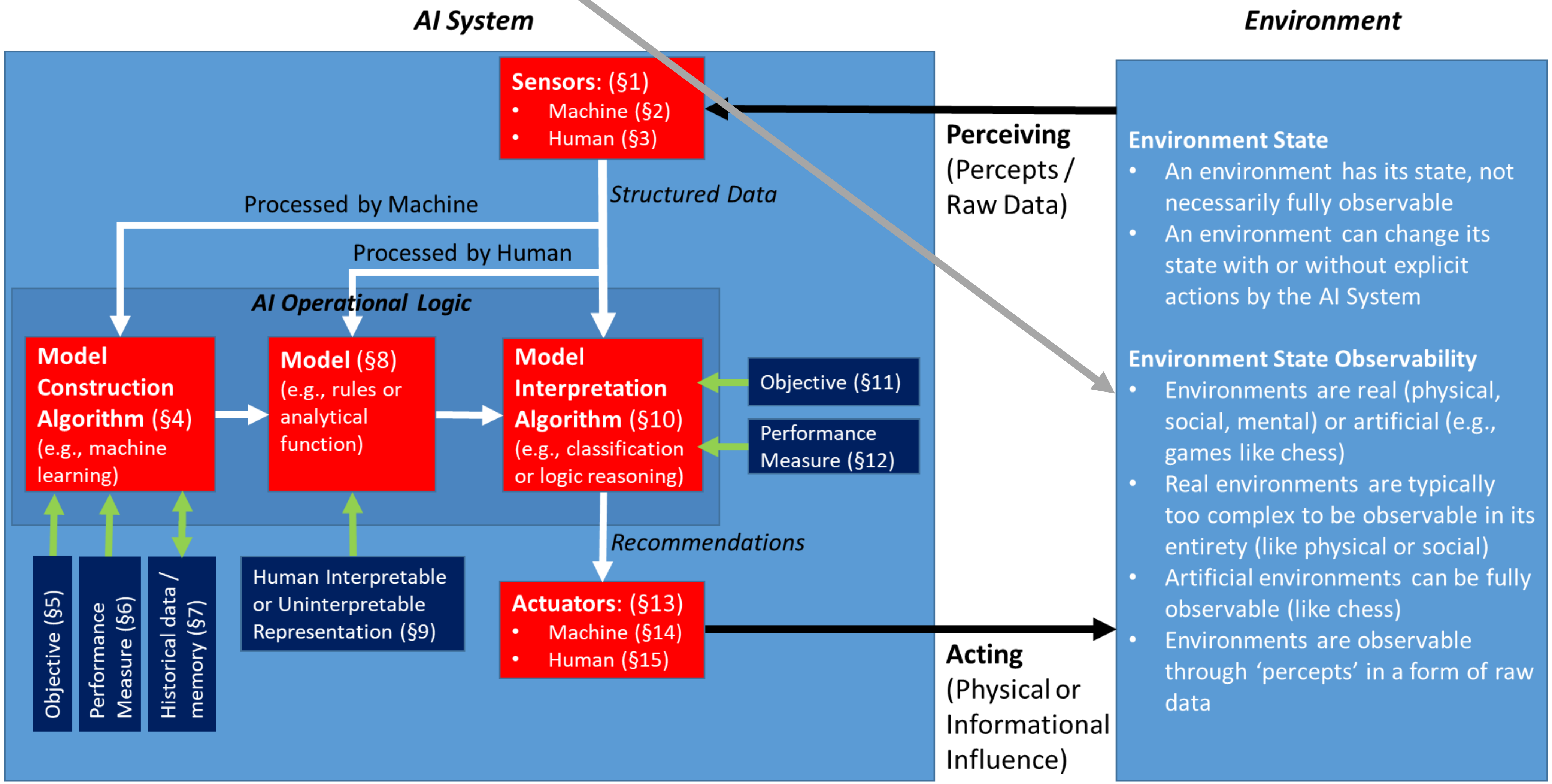
Acting (Physical or Informational Influence)

Environment

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through 'percepts' in a form of raw data

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



AI System

Sensors: (§1)
- Machine (§2)
- Human (§3)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

Perceiving (Percepts / Raw Data)

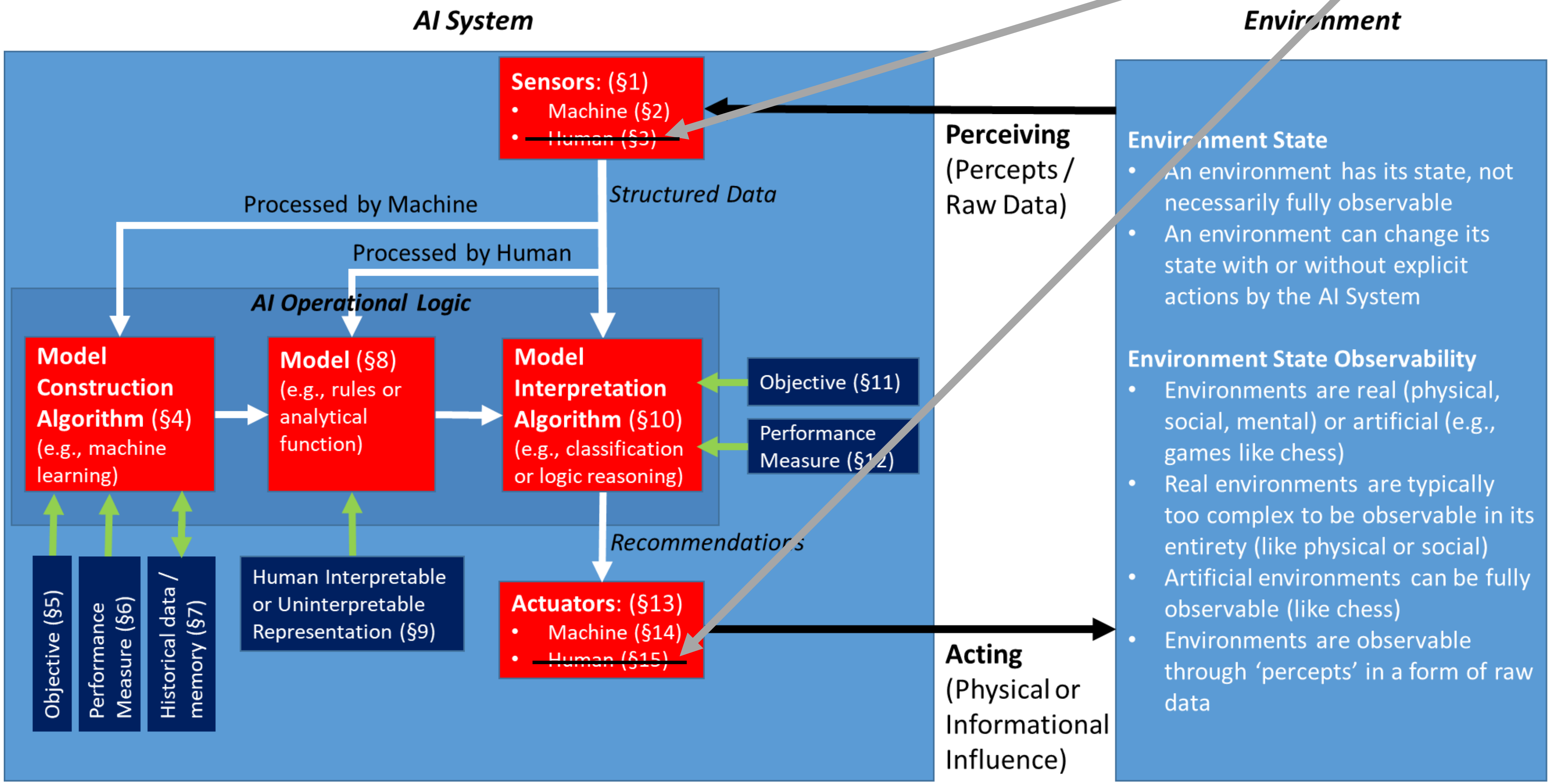
Acting (Physical or Informational Influence)

Environment

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through 'percepts' in a form of raw data

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



AI System

Environment

Sensors: (§1)
- Machine (§2)
- ~~Human (§3)~~

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- ~~Human (§15)~~

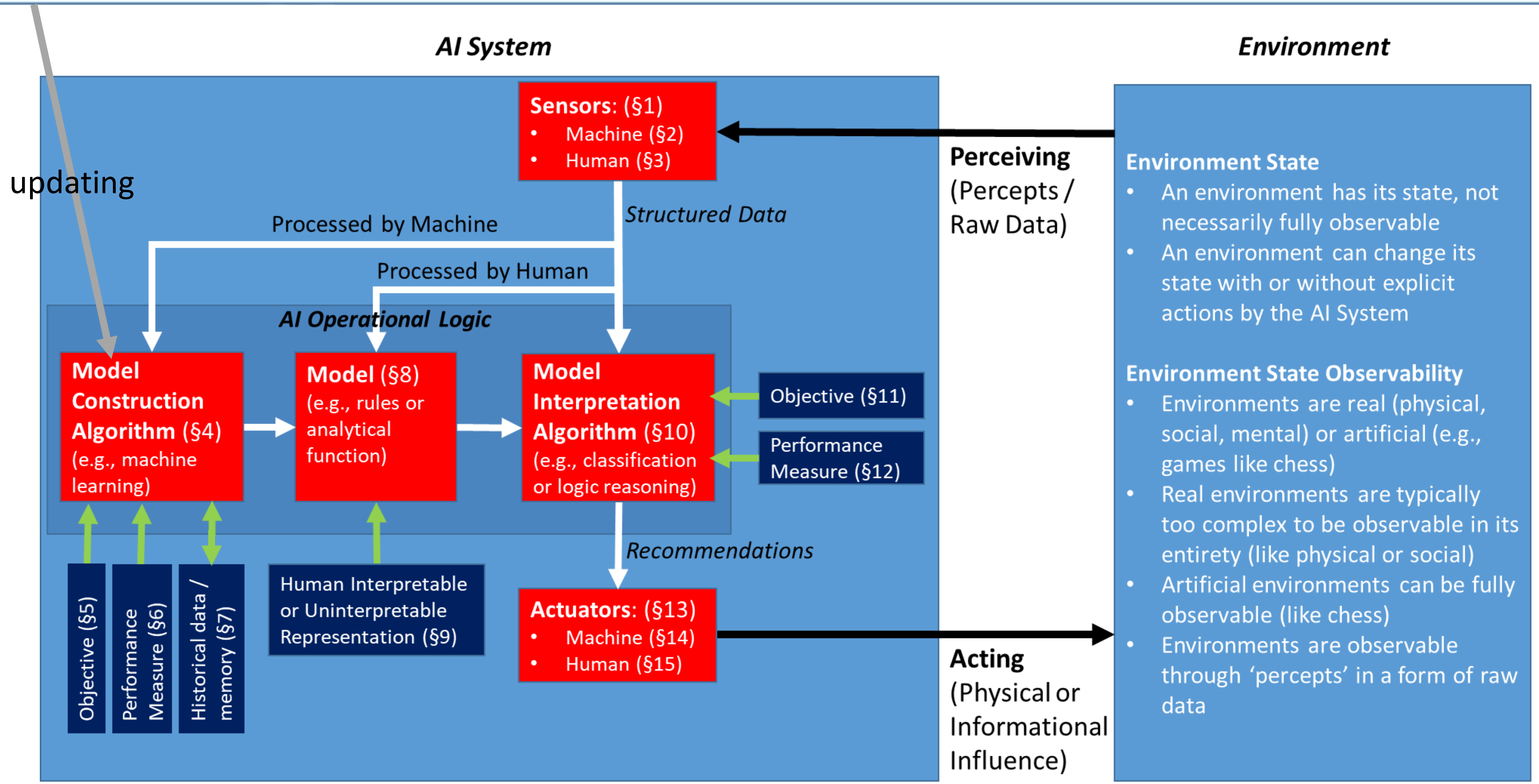
Perceiving (Percepts / Raw Data)

Acting (Physical or Informational Influence)

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through 'percepts' in a form of raw data

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



Sensors: (§1)
- Machine (§2)
- Human (§3)

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

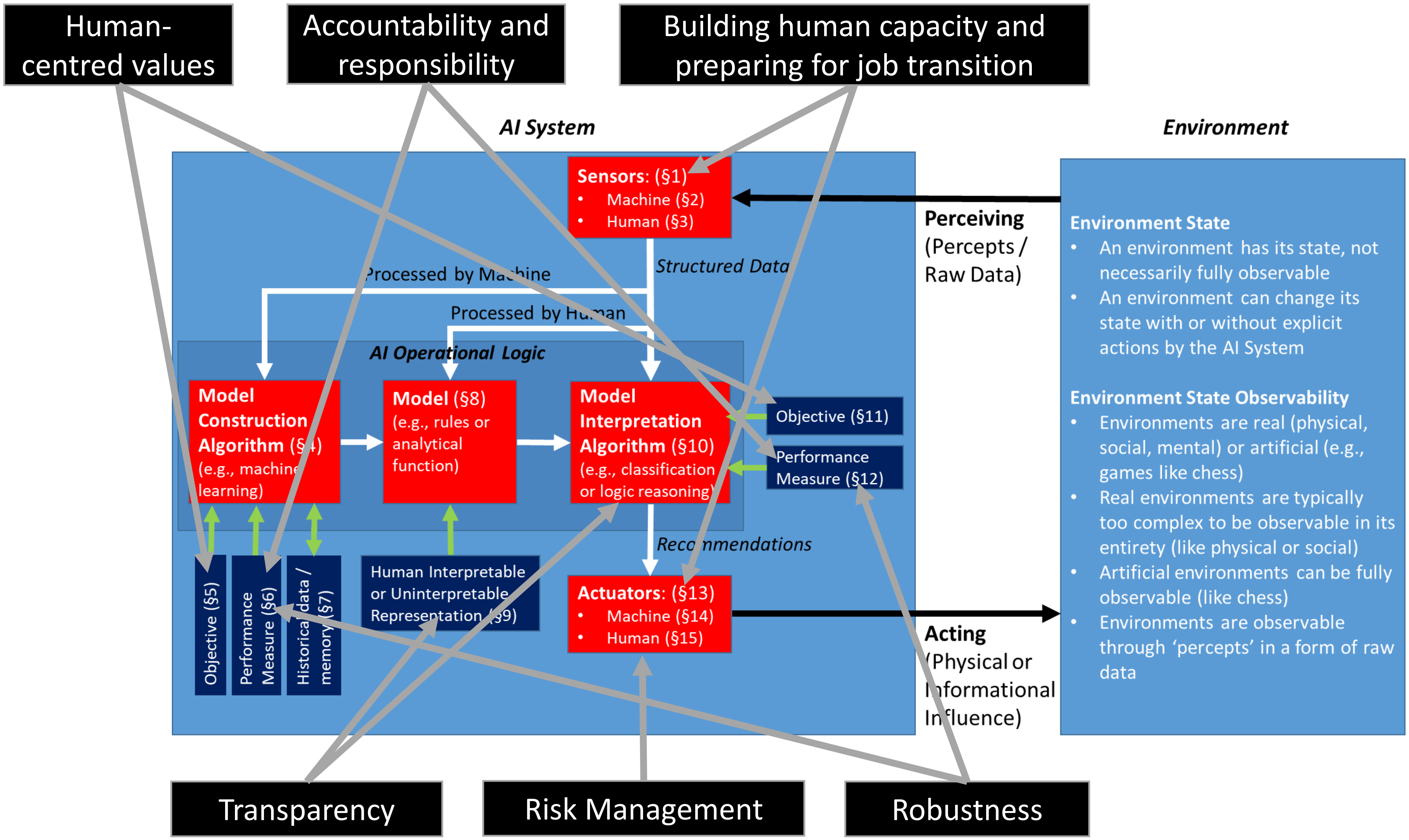
Human Interpretable or Uninterpretable Representation (§9)

Actuators: (§13)
- Machine (§14)
- Human (§15)

Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

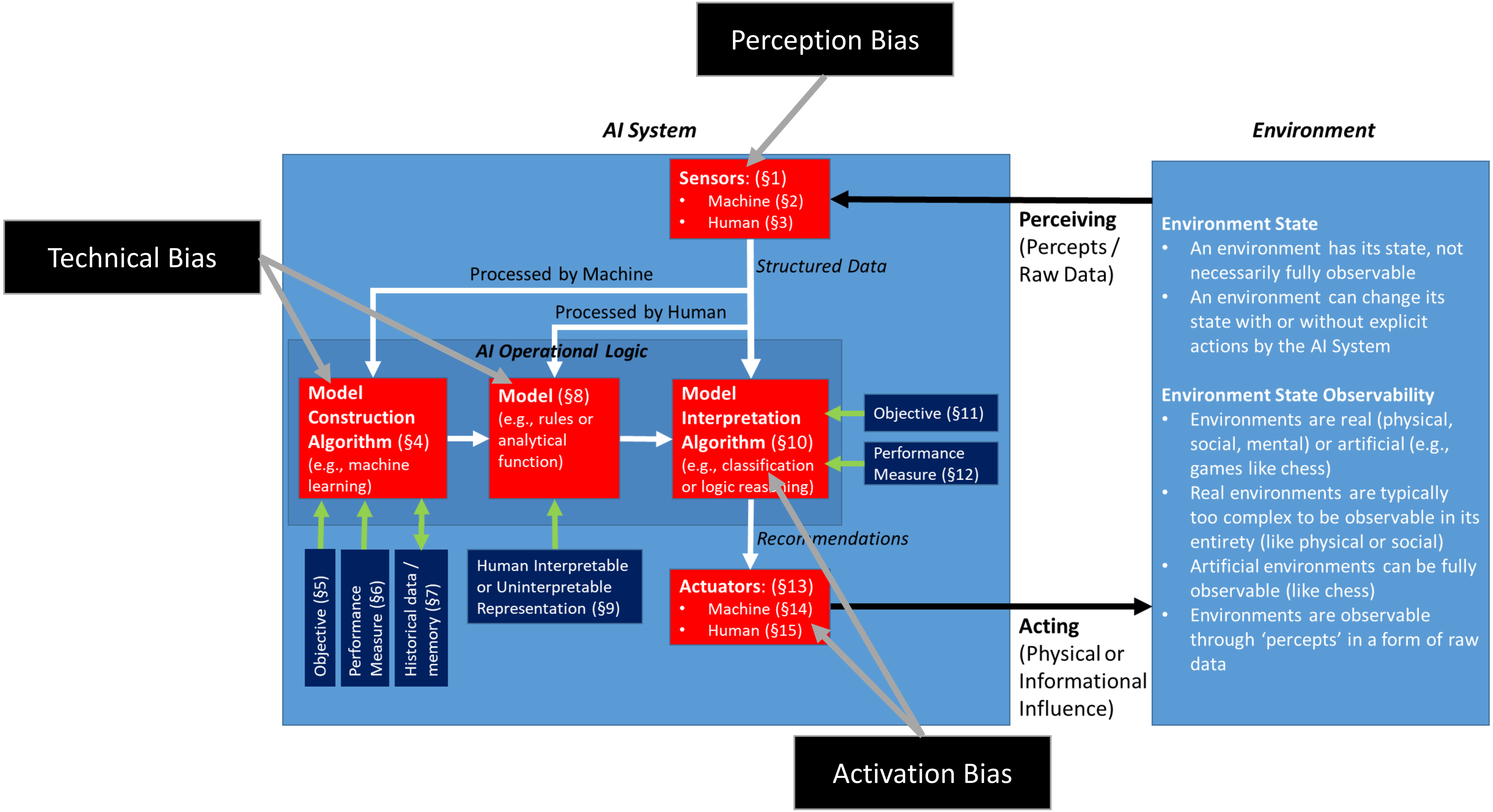
Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through 'percepts' in a form of raw data

AI System and relation to higher level principles



Human-centred values
Accountability and responsibility
Building human capacity and preparing for job transition
AI System
Environment
Sensors: (§1)
• Machine (§2)
• Human (§3)
Structured Data
Processed by Machine
Processed by Human
AI Operational Logic
Model Construction Algorithm (§4) (e.g., machine learning)
Model (§8) (e.g., rules or analytical function)
Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)
Objective (§11)
Performance Measure (§12)
Objective (§5)
Performance Measure (§6)
Historical data / memory (§7)
Human Interpretable or Uninterpretable Representation (§9)
Recommendations
Actuators: (§13)
• Machine (§14)
• Human (§15)
Perceiving (Percepts / Raw Data)
Acting (Physical or Informational Influence)
Environment State
• An environment has its state, not necessarily fully observable
• An environment can change its state with or without explicit actions by the AI System
Environment State Observability
• Environments are real (physical, social, mental) or artificial (e.g., games like chess)
• Real environments are typically too complex to be observable in its entirety (like physical or social)
• Artificial environments can be fully observable (like chess)
• Environments are observable through 'percepts' in a form of raw data
Transparency
Risk Management
Robustness

AI System and sources of various types of biases



Perception Bias

Technical Bias

Activation Bias

AI System

Environment

Sensors: (§1)
- Machine (§2)
- Human (§3)

Structured Data

Processed by Machine

Processed by Human

AI Operational Logic

Model Construction Algorithm (§4) (e.g., machine learning)

Model (§8) (e.g., rules or analytical function)

Model Interpretation Algorithm (§10) (e.g., classification or logic reasoning)

Objective (§11)

Performance Measure (§12)

Objective (§5)

Performance Measure (§6)

Historical data / memory (§7)

Human Interpretable or Uninterpretable Representation (§9)

Recommendations

Actuators: (§13)
- Machine (§14)
- Human (§15)

Perceiving (Percepts / Raw Data)

Acting (Physical or Informational Influence)

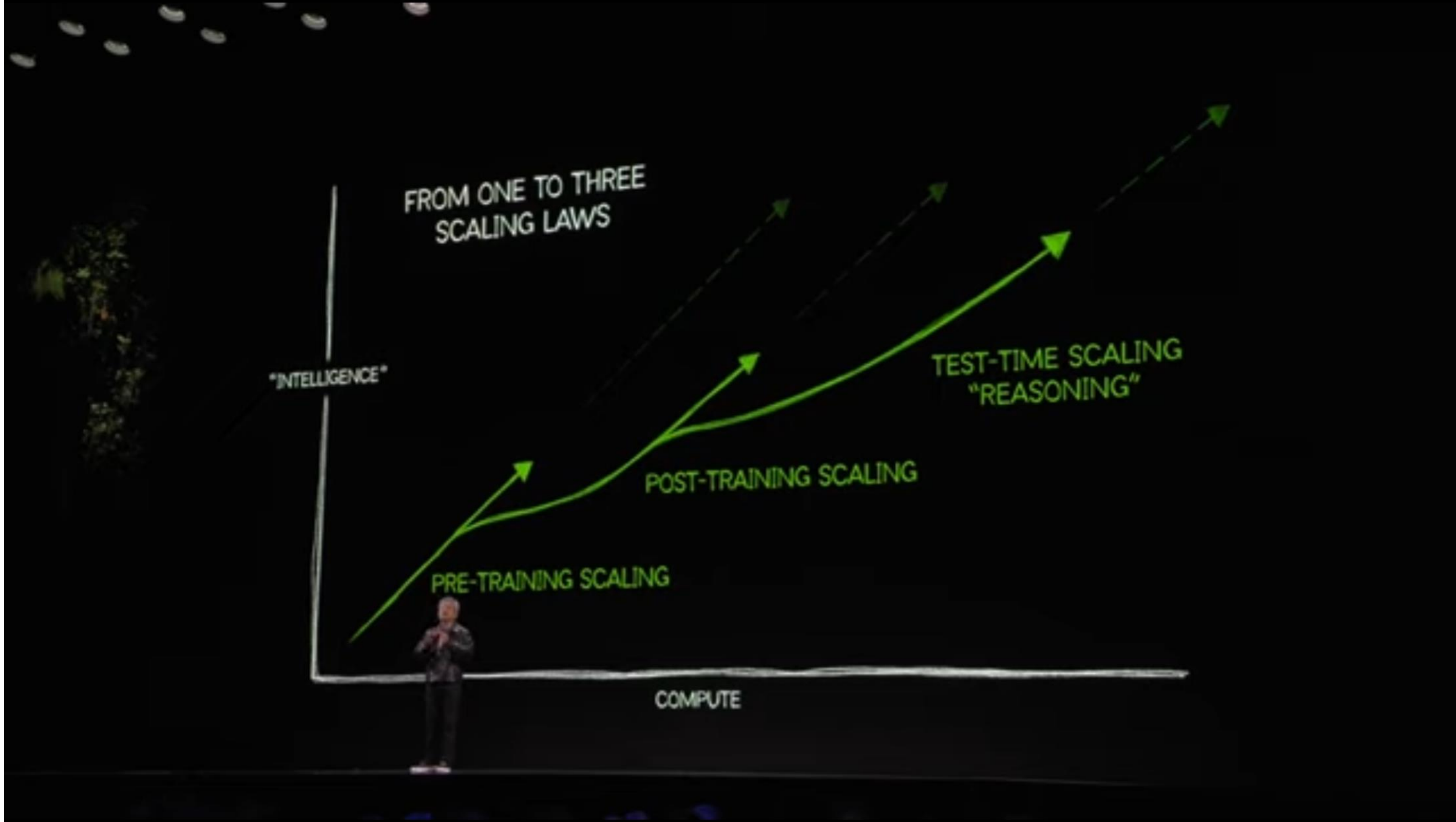
Environment State
- An environment has its state, not necessarily fully observable
- An environment can change its state with or without explicit actions by the AI System

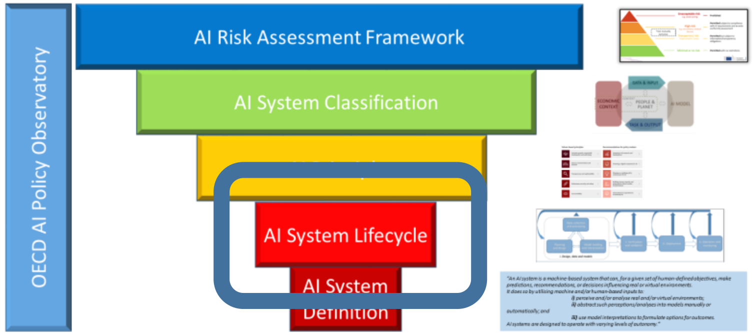
Environment State Observability
- Environments are real (physical, social, mental) or artificial (e.g., games like chess)
- Real environments are typically too complex to be observable in its entirety (like physical or social)
- Artificial environments can be fully observable (like chess)
- Environments are observable through ‘percepts’ in a form of raw data

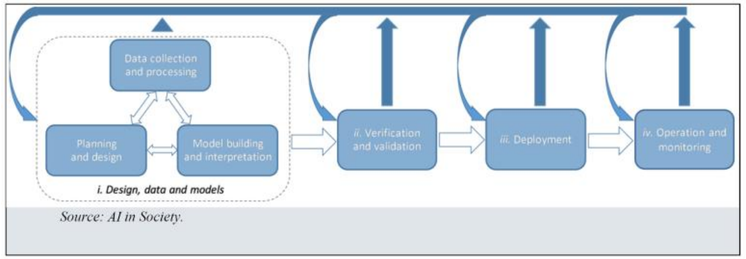
Three Levels of AI Scaling

- Pre-Training, Post-Trainig, and Test-Time Scaling “Reasoning”

- “Reasoning” becoming in 2025 as likely the main topic





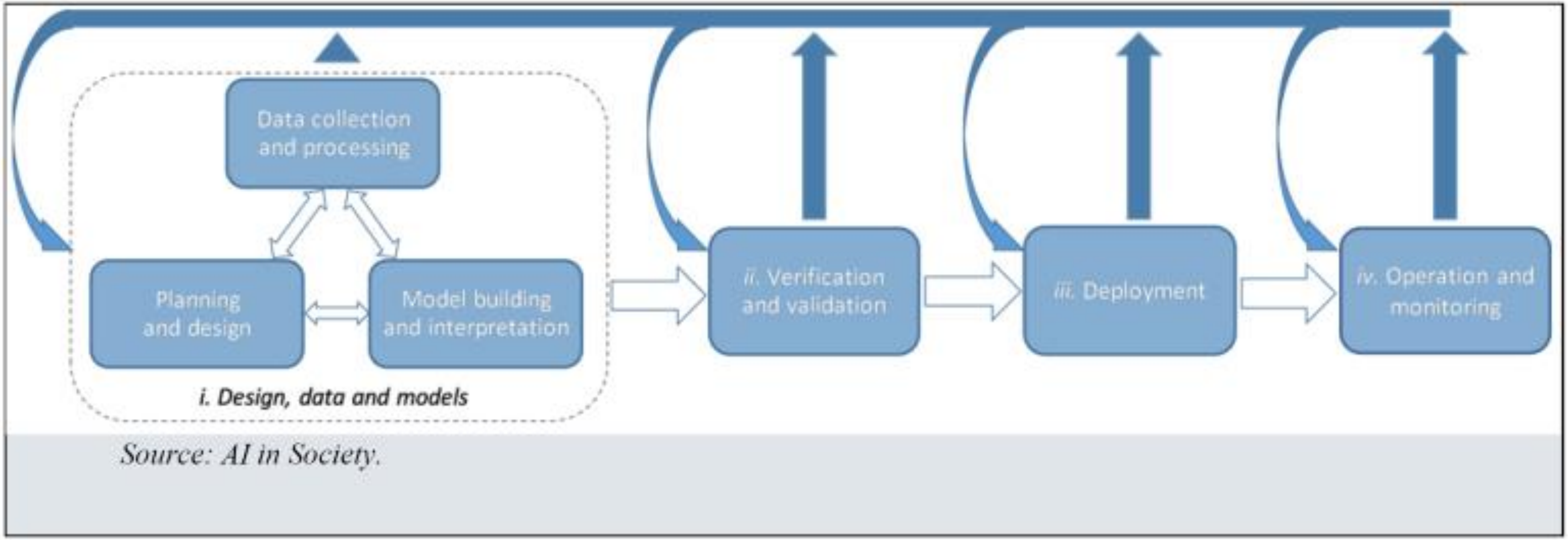


Source: AI in Society.

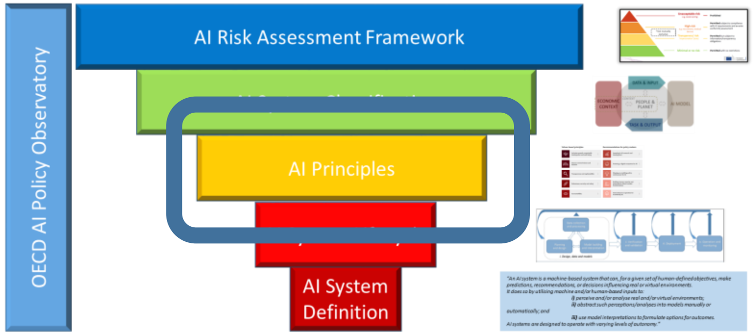
OECD AI System Lifecycle

Adopted in 2019

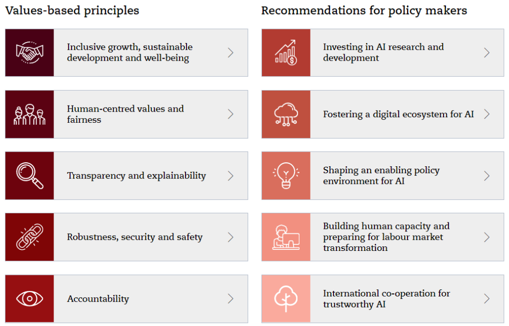
OECD AI System Lifecycle


Data collection and processing
Planning and design
Model building and interpretation
i. Design, data and models
ii. Verification and validation
iii. Deployment
iv. Operation and monitoring

Source: AI in Society.



"An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. It does so by utilising machine and/or human-based inputs to:
i) perceive and/or analyse real and/or virtual environments;
ii) abstract such perceptions/analyses into models manually or automatically; and
iii) use model interpretations to formulate options for outcomes.
AI systems are designed to operate with varying levels of autonomy."



Values-based principles	Recommendations for policy makers
Inclusive growth, sustainable development and well-being	Investing in AI research and development
Human-centred values and fairness	Fostering a digital ecosystem for AI
Transparency and explainability	Shaping an enabling policy environment for AI
Robustness, security and safety	Building human capacity and preparing for labour market transformation
Accountability	International co-operation for trustworthy AI

OECD AI Principles

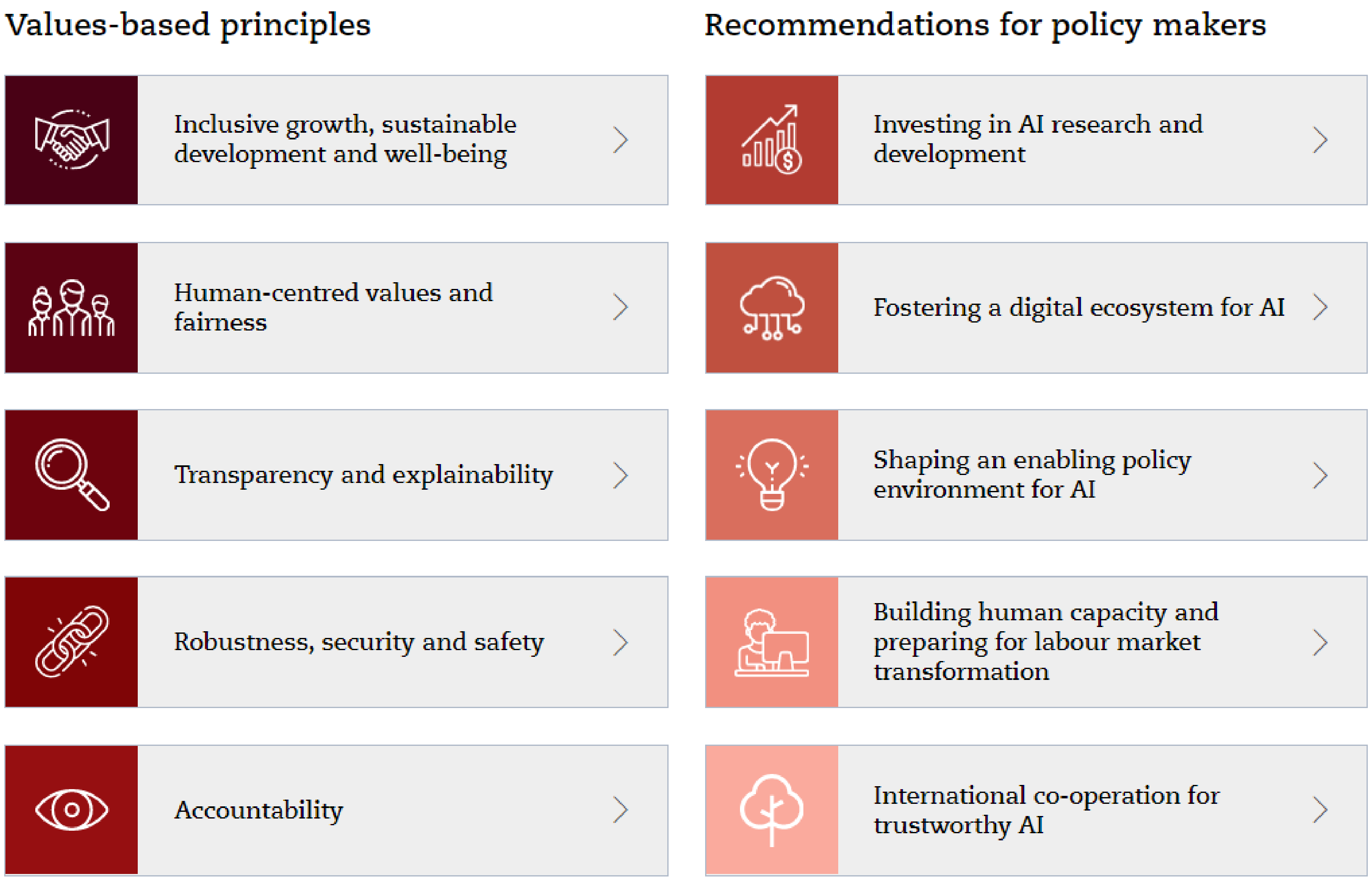
Adopted in 2019

https://oecd.ai/en/ai-principles

OECD AI Principles

(the only politically agreed AI document so far – 44 countries)

https://oecd.ai/ai-principles



Values-based principles

- Inclusive growth, sustainable development and well-being
- Human-centred values and fairness
- Transparency and explainability
- Robustness, security and safety
- Accountability

Recommendations for policy makers

- Investing in AI research and development
- Fostering a digital ecosystem for AI
- Shaping an enabling policy environment for AI
- Building human capacity and preparing for labour market transformation
- International co-operation for trustworthy AI

Legal document: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

1.1.Inclusive growth, sustainable development and well-being

Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

1.2.Human-centred values and fairness

a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.

b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

1.3.Transparency and explainability

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

i. to foster a general understanding of AI systems,

ii. to make stakeholders aware of their interactions with AI systems, including in the workplace,

iii. to enable those affected by an AI system to understand the outcome, and,

iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

1.4.Robustness, security and safety

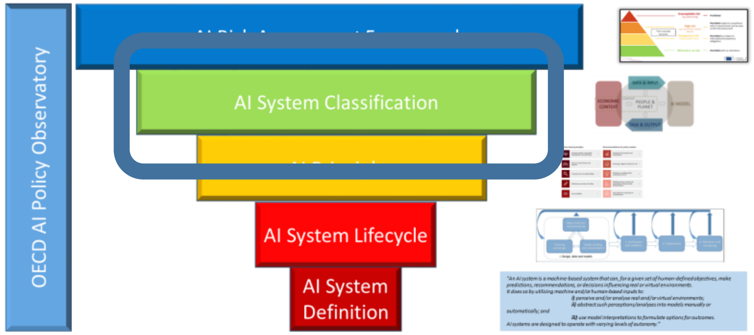
a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.

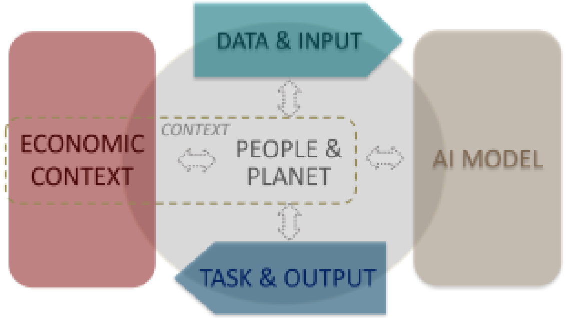
b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system’s outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

c) 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.

1.5.Accountability

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





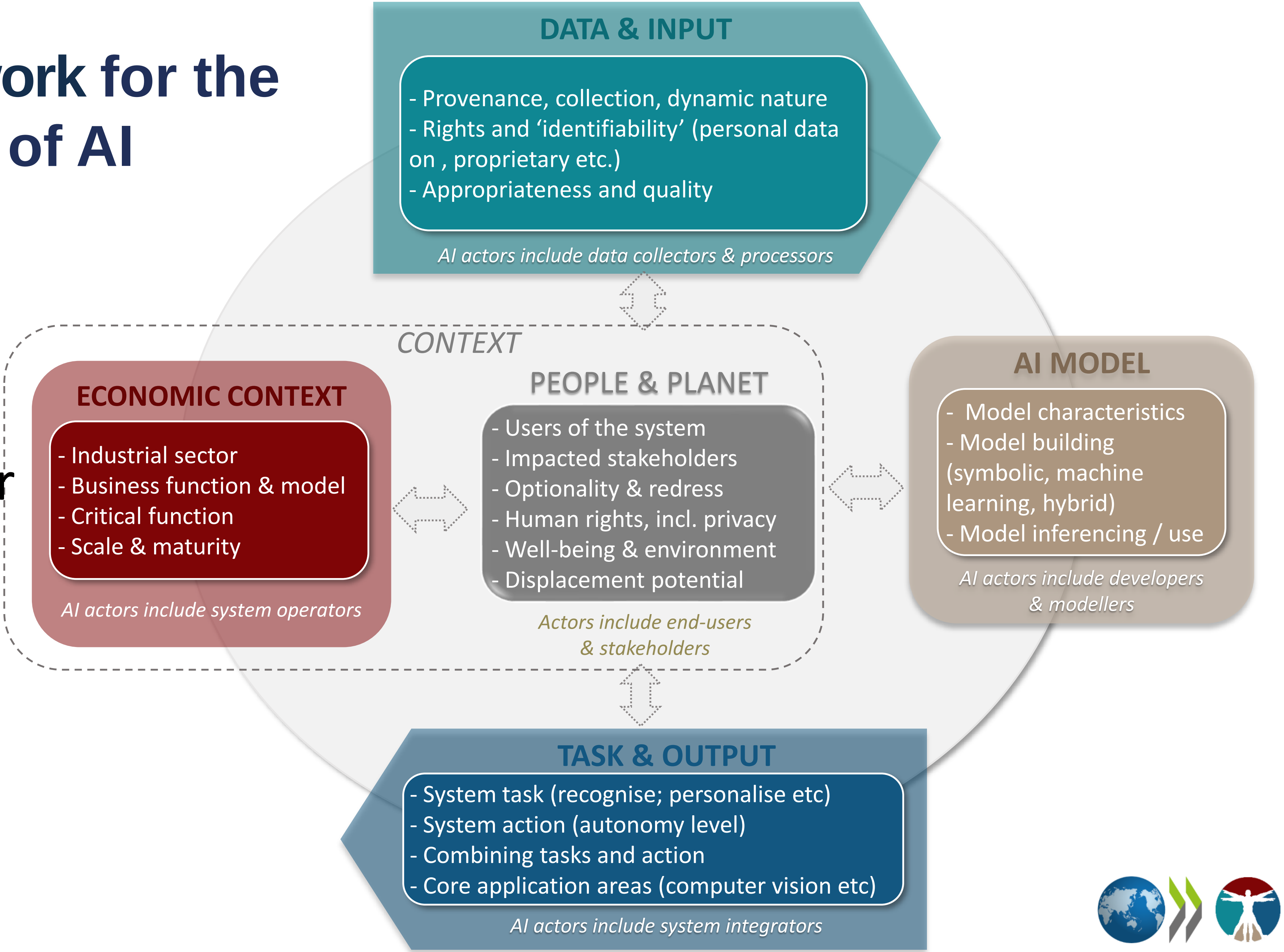
OECD AI System Classification

Adopted in 2022

https://oecd.ai/en/classification

OECD framework for the classification of AI systems

Top level dimensions include a number of sub-dimensions equipped with measurable indicators



DATA & INPUT

- Provenance, collection, dynamic nature
- Rights and ‘identifiability’ (personal data on , proprietary etc.)
- Appropriateness and quality

AI actors include data collectors & processors

CONTEXT

ECONOMIC CONTEXT

- Industrial sector
- Business function & model
- Critical function
- Scale & maturity

AI actors include system operators

PEOPLE & PLANET

- Users of the system
- Impacted stakeholders
- Optionality & redress
- Human rights, incl. privacy
- Well-being & environment
- Displacement potential

Actors include end-users & stakeholders

AI MODEL

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

AI actors include developers & modellers

TASK & OUTPUT

- System task (recognise; personalise etc)
- System action (autonomy level)
- Combining tasks and action
- Core application areas (computer vision etc)

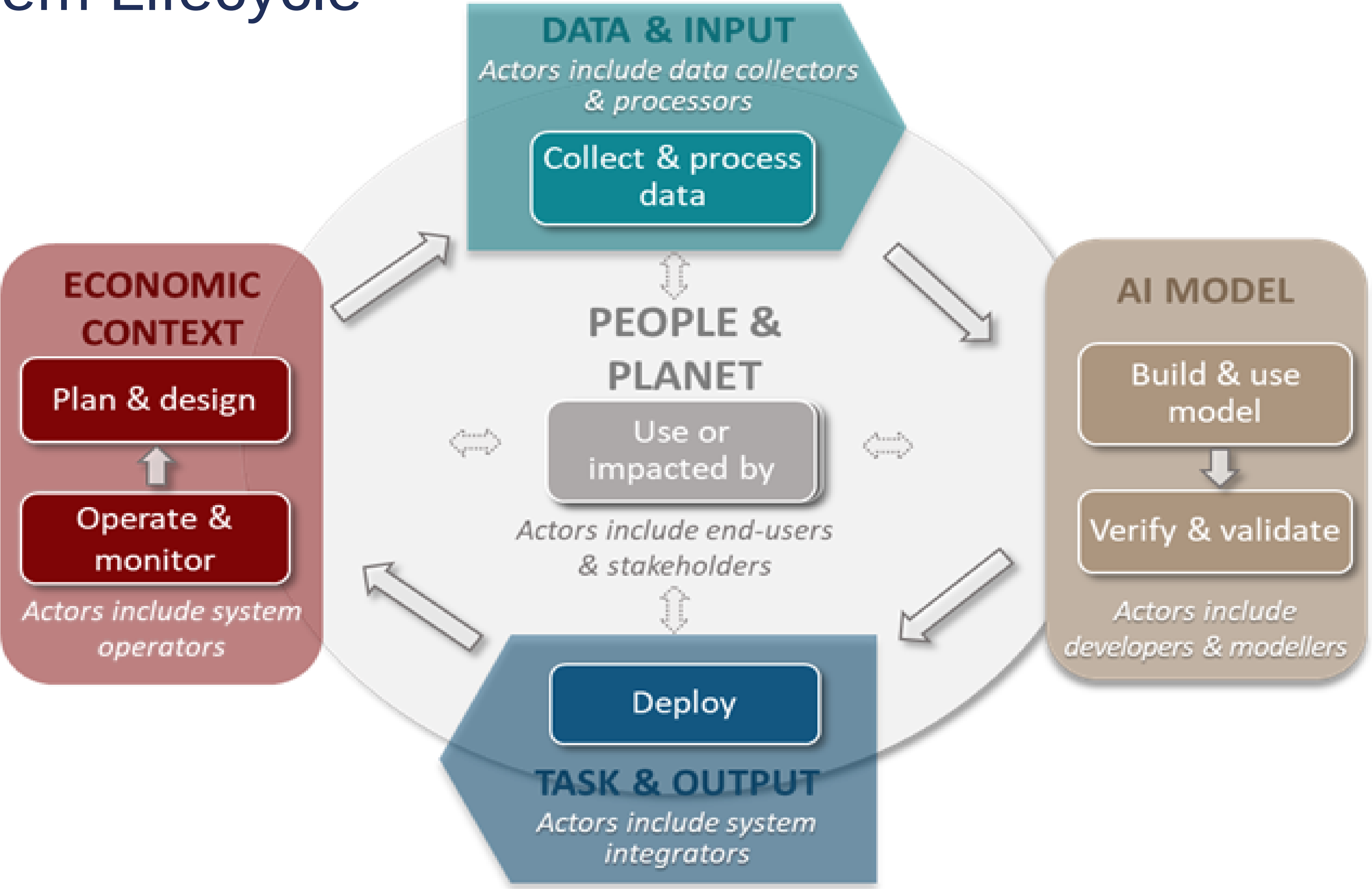
AI actors include system integrators

Linking the classification & AI system lifecycle actors

Framework dimensions	*People & Planet*	Economic Context		Data & Input	AI Model		Task & Output
Actors include	*End-users & stakeholders*	System operators		Data collectors & processors	Developers & modellers		System integrators
Lifecycle stage	*Use or are impact by*	Plan & design	Operate & monitor	Collect & process data	Build & use	Build & validate	Deploy

OECD framework for the classification of AI systems

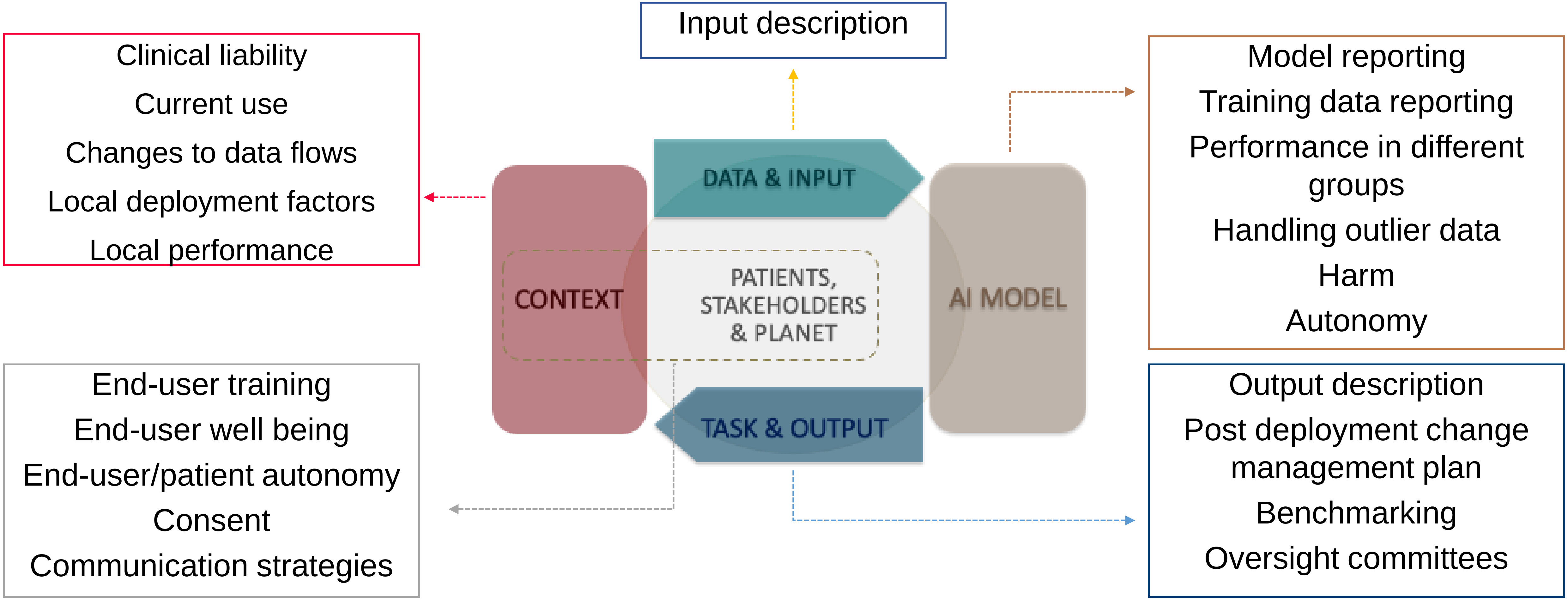
AI System Lifecycle


DATA & INPUT
Actors include data collectors & processors
Collect & process data
AI MODEL
Build & use model
Verify & validate
Actors include developers & modellers
Deploy
TASK & OUTPUT
Actors include system integrators
ECONOMIC CONTEXT
Plan & design
Operate & monitor
Actors include system operators
PEOPLE & PLANET
Use or impacted by
Actors include end-users & stakeholders

Using the framework for health technology assessment

The Alan Turing Institute

UNIVERSITY OF BIRMINGHAM



Input description

Clinical liability

Current use

Changes to data flows

Local deployment factors

Local performance

Model reporting

Training data reporting

Performance in different groups

Handling outlier data

Harm

Autonomy

End-user training

End-user well being

End-user/patient autonomy

Consent

Communication strategies

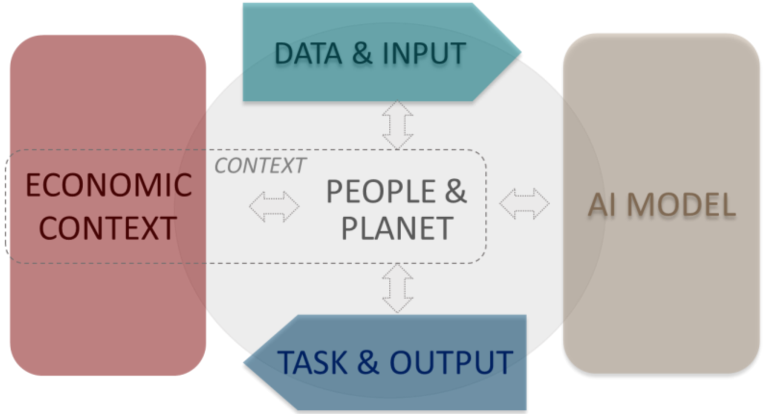
Output description

Post deployment change management plan

Benchmarking

Oversight committees

Example 1: Credit-scoring AI systems



Selected criteria:

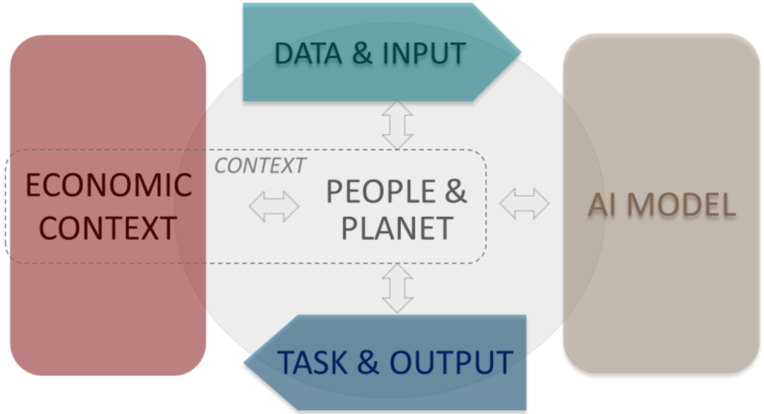
- **System users –** Amateur (bank employee)
- **Optionality –** Cannot opt out
- **Human rights impact –** Yes
- **Sector of deployment** – **Financial system (e.g., banking, insurance)**
- **Critical function** – **Critical function/activity (availability of financial services, inclusion)**
- **Data collection –** Human (set of rules) and automated sources (e.g. profiles, loan payments)
- **Rights** – Mix of proprietary and public data
- **"Identifiability" –** often personally identifiable data
- **Model building** – e.g., statistical/hybrid model; learns from provided data, augmented by human knowledge
- **Model evolution** – Can evolve during operation
- **System task** – Forecasting: uses past & existing behavior to predict future outcomes
- **Level of action autonomy** – Medium (human on-the-loop)

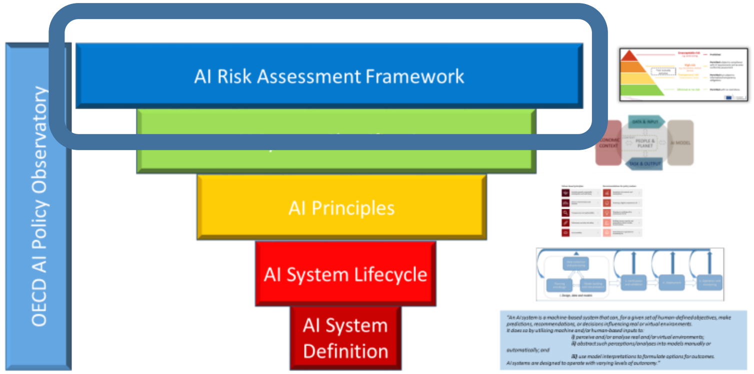
Example 2: GPT-3, text generation

Selected criteria:

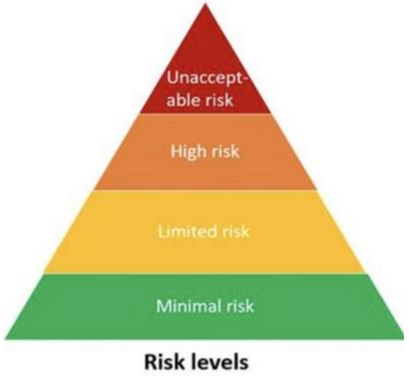
Caveat: general purpose AI system, so nearly all responses depend on the specific application context! Medical advice, content filter, creative writing…

- **System users** – Primary users are amateur
- **Impacted stakeholders** – workers, consumers
- **Sector of deployment** – Information & communication
- **Critical function** – None
- **Data collection** – Human sources (text strings)
- **Rights** – Largely public data sources (some proprietary)
- **Model building** – Learn from provided data
- **Model evolution** – Evolution during operation
- **System task** – Goal-driven optimization, Reasoning with knowledge structures, interaction support, recognition, personalisation
- **Level of action autonomy** – Low autonomy [human action required e.g., to use generated text]





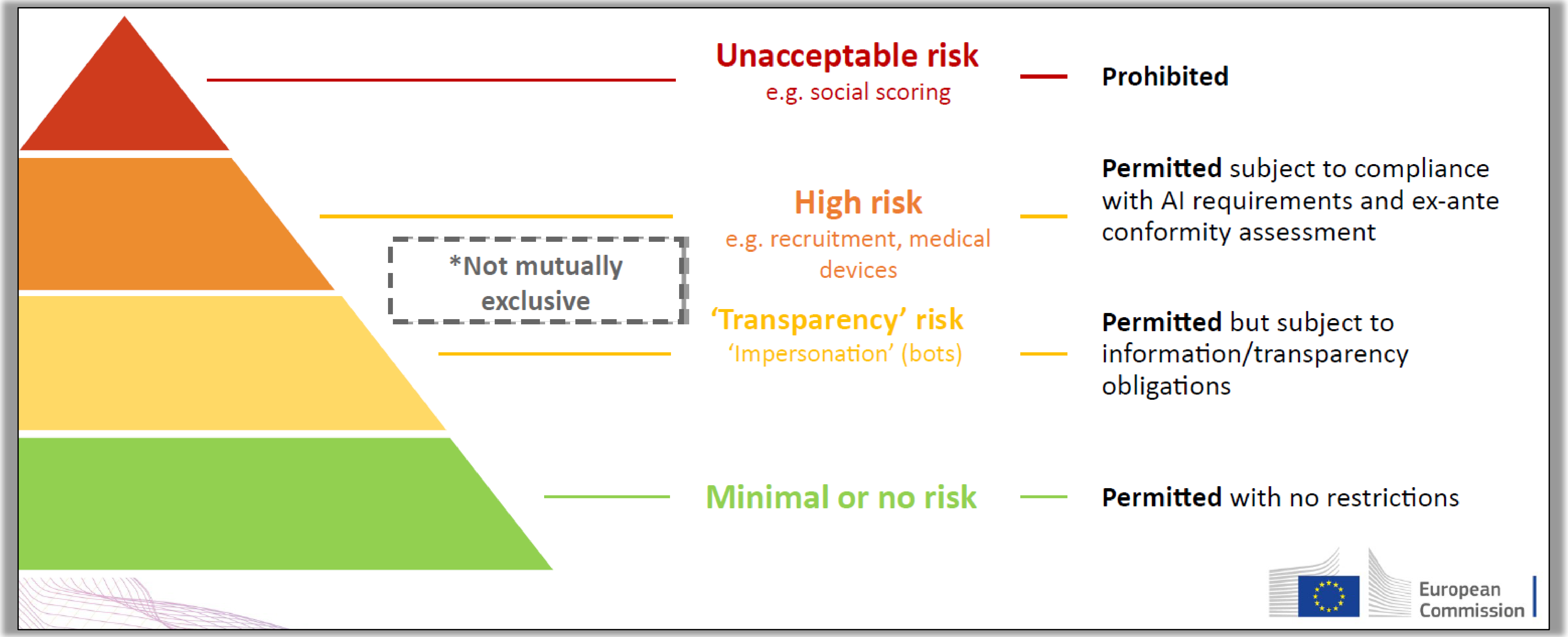
"An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.
It does so by utilising machine and/or human-based inputs to:
i) perceive and/or analyse real and/or virtual environments;
ii) abstract such perceptions/analyses into models manually or automatically; and
iii) use model interpretations to formulate options for outcomes.
AI systems are designed to operate with varying levels of autonomy."



OECD AI Risk Assessment

...work in progress

OECD Risk assessment framework: categorization of **uses of AI** in the draft EU AI Act



Unacceptable risk
e.g. social scoring
— **Prohibited**

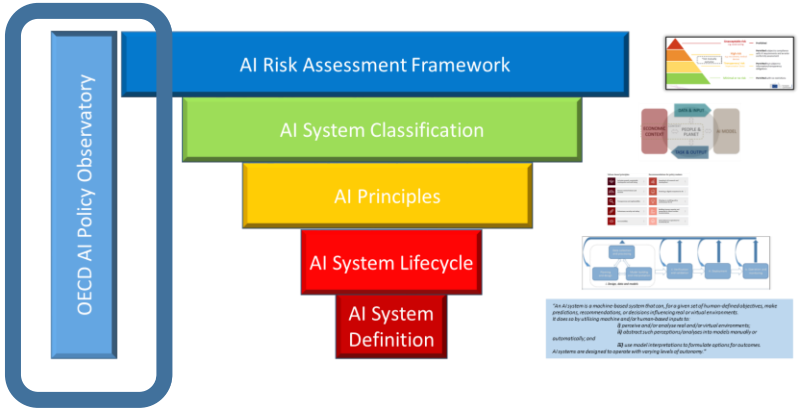
High risk
e.g. recruitment, medical devices
— **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

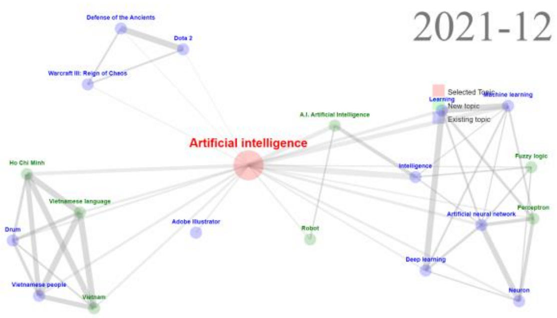
*Not mutually exclusive

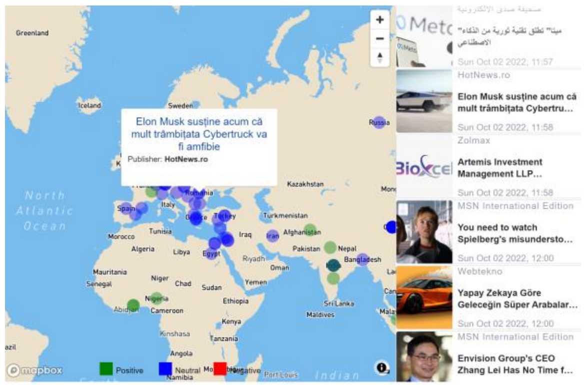
'Transparency' risk
'Impersonation' (bots)
— **Permitted** but subject to information/transparency obligations

Minimal or no risk
— **Permitted** with no restrictions

European Commission







Elon Musk susține acum că mult trâmbițata Cybertruck va fi amfibie
Publisher: HotNews.ro

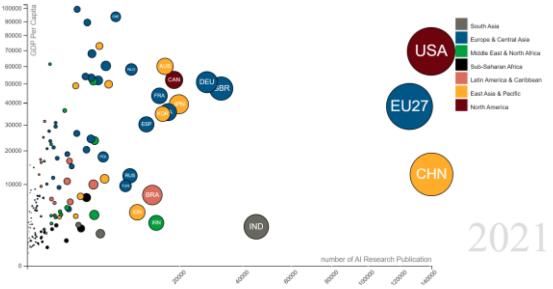
Sun Oct 02 2022, 11:57
HotNews.ro
Elon Musk susține acum că mult trâmbițata Cybertru...

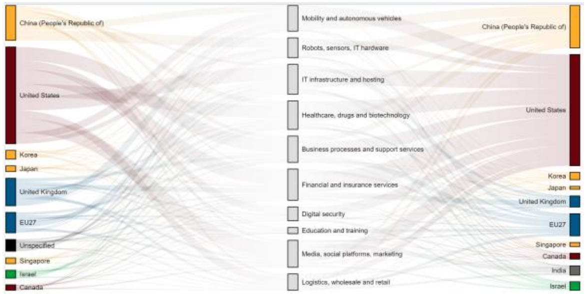
Sun Oct 02 2022, 11:58
Zolmax
Artemis Investment Management LLP...

Sun Oct 02 2022, 11:58
MSN International Edition
You need to watch Spielberg's misundersto...

Sun Oct 02 2022, 12:00
Webtekno
Yapay Zekaya Göre Geleceğin Süper Arabalar...

Sun Oct 02 2022, 12:00
MSN International Edition
Envision Group's CEO Zhang Lei Has No Time f...



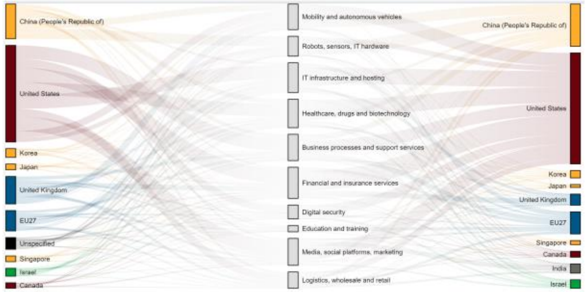


OECD AI Policy Observatory

Near real-time observation of the evolution of AI across 12 dimensions

https://oecd.ai/

Real-Time Technology Watch *“a journey of an innovation”*



- “OECD AI Policy Observatory” (https://oecd.ai/)

- Main objectives of the use case are to build a platform to respond on questions related to the global innovation ecosystem in the area of AI
    - *To understand the evolution of AI?*
    - *To detect impactful innovations early in the process?*
    - *To predict what will be ‘the next big thing’ in AI?*
    - *Building aka ‘the digital twin of AI ecosystem’*

- The basic premise is that ideas and innovations which will impact our lives in the next 5-10 years are already invented and published…

OECD AI Policy Observatory narrative:

Tracking an innovation across many stages of the ecosystem

- An innovation **spotted in the academic world**…
- …**projects** are started around the innovation (publicly funded, open source)
- …researchers & developers **informally discuss** the innovation
- …the innovation gets **patented**
- …**companies** are established around the innovation
- …companies get **investments**, possibly in several rounds
- …investments have influence on **job market** (supply and demand side)
- …**market** reacts on the quality of innovation
- …**education** introduces new courses
- …**perception** & interest from expert and broad audiences
- …**media** starts publishing about the innovation and companies
- …**incidents** happen to show weaknesses to be treated
- …**policies** are formulated on international and national level

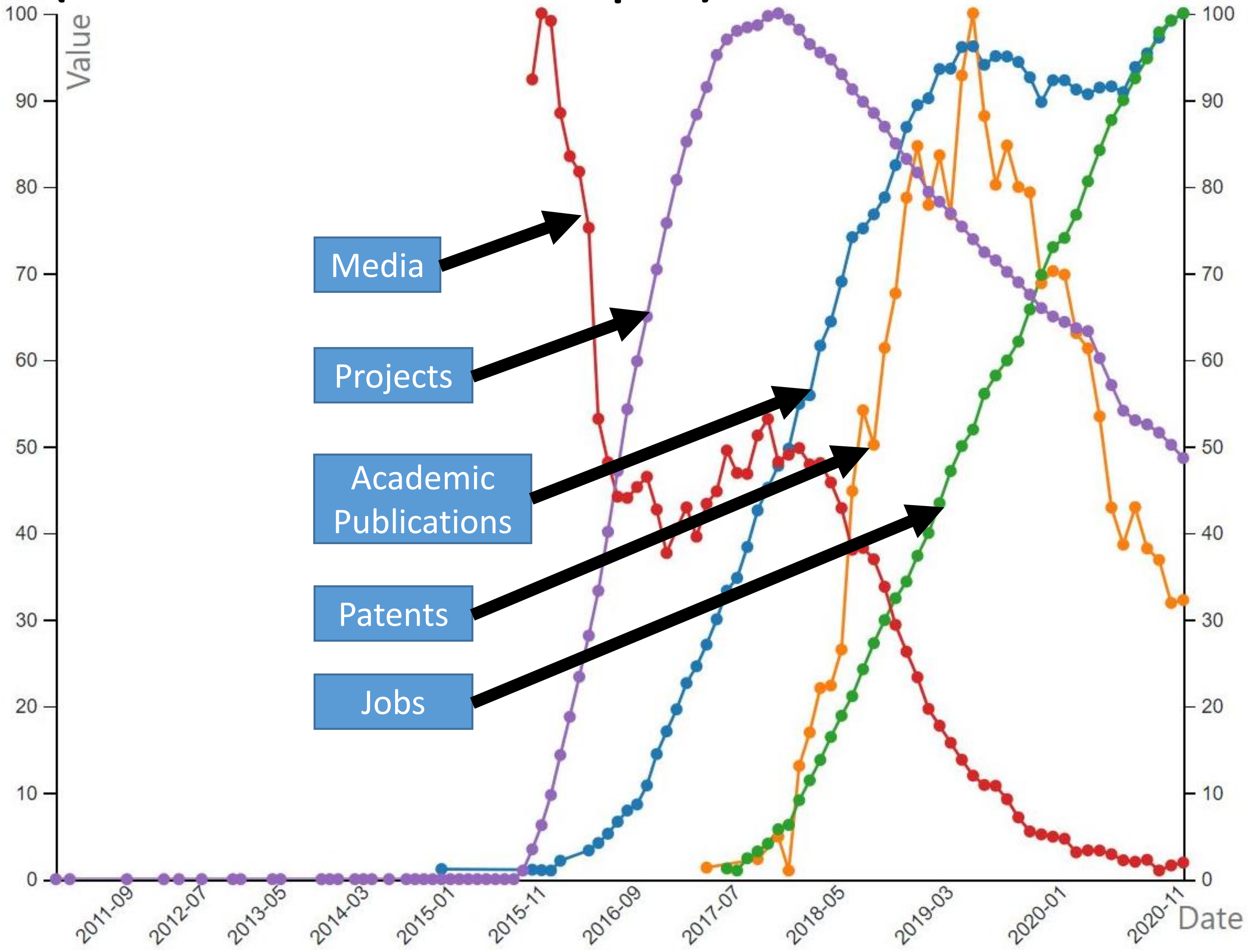
OECD AI Policy Observatory (oecd.ai) data sources

- **Academic world** – Microsoft Academic Graph/OpenAlex, SCOPUS (~200M, ~1M per month)
- **Projects** – CORDIS/NSF/… (>100k), GitHub (~30M repositories)
- **Informally discussions** – StackOverflow.com forums
- **Patents** – Microsoft Academic Graph
- **Companies** – Orbis, Dun & Bradstreet
- **Investments** – Preqin.com (>20k investments)
- **Job market –** LinkedIn.com (supply side) and Adzuna.com (demand side)
- **Market** – Yahoo Finance, Bloomberg, …
- **Education** – StudyPortals.com (~3000 universities, English courses only)
- **Perception** – Google Trends & Twitter
- **Media** – EventRegistry.org (1M news per day)
- **Incidents** – database in construction (>1000) based on IncidentDatabase.ai
- **Policies** – OECD global policies database (oecd.ai) (~1000 docs on AI)

Cascading influence of an innovation (“tensorflow” example)

- Impact of an innovation to the ecosystem
- Example for “**Google TensorFlow**” used by all of us many times per day
- Cascading influence:
  - Starting with **media**,
  - …triggering **projects**,
  - …resulting in **academic publications**,
  - …followed by **patents**,
  - …influencing **job market**

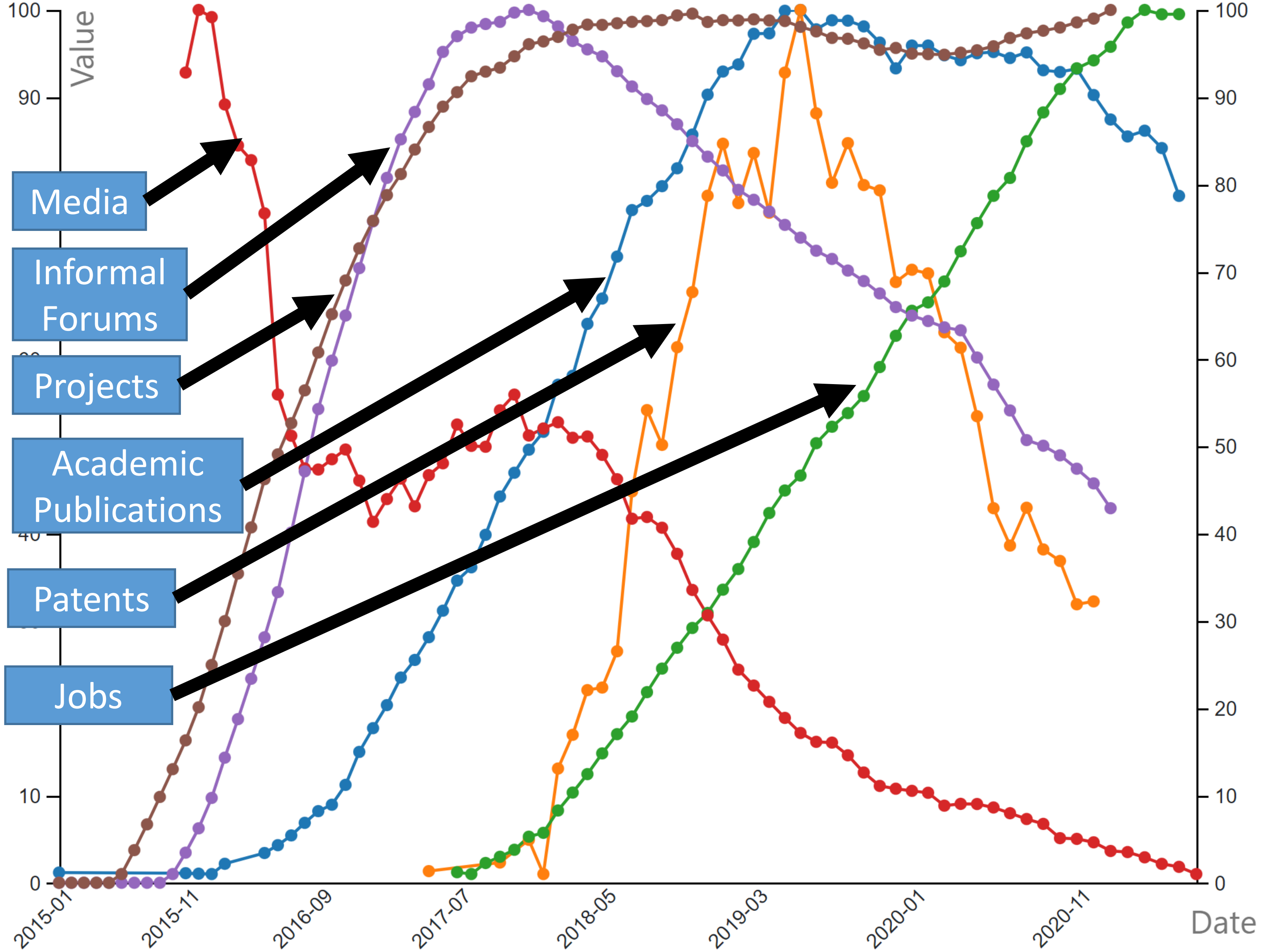
https://aibench.ijs.si/



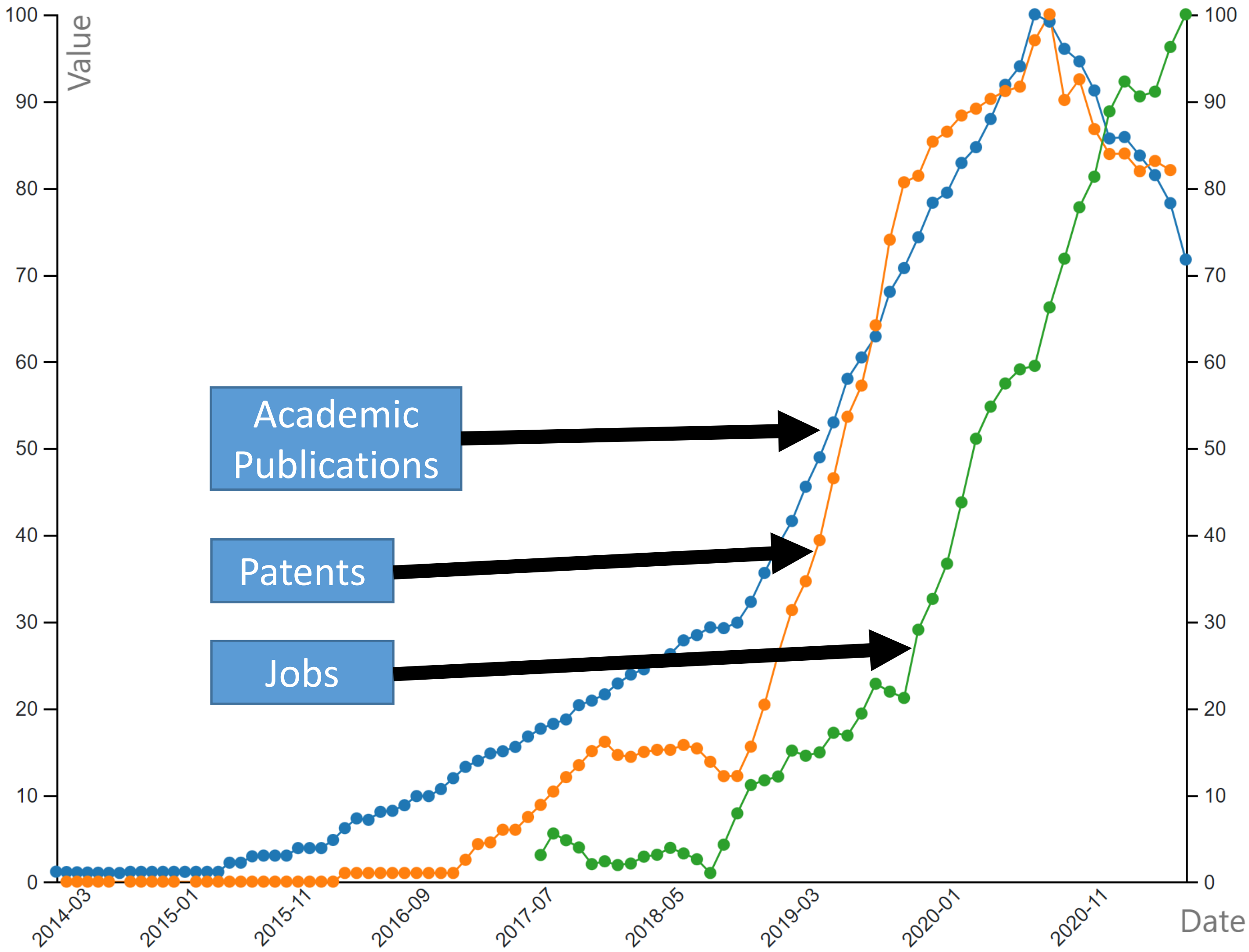
Cascading influence of an innovation ("tensorflow" example)

- Impact of an innovation to the ecosystem
- Example for "**Google TensorFlow**" used by all of us many times per day
- Cascading influence:
  - Starting with **media**,
  - …triggering **projects**,
  - …resulting in **academic publications**,
  - …followed by **patents**,
  - …influencing **job market**

https://aibench.ijs.si/



Cascading influence of an innovation (“knowledge graph” example)

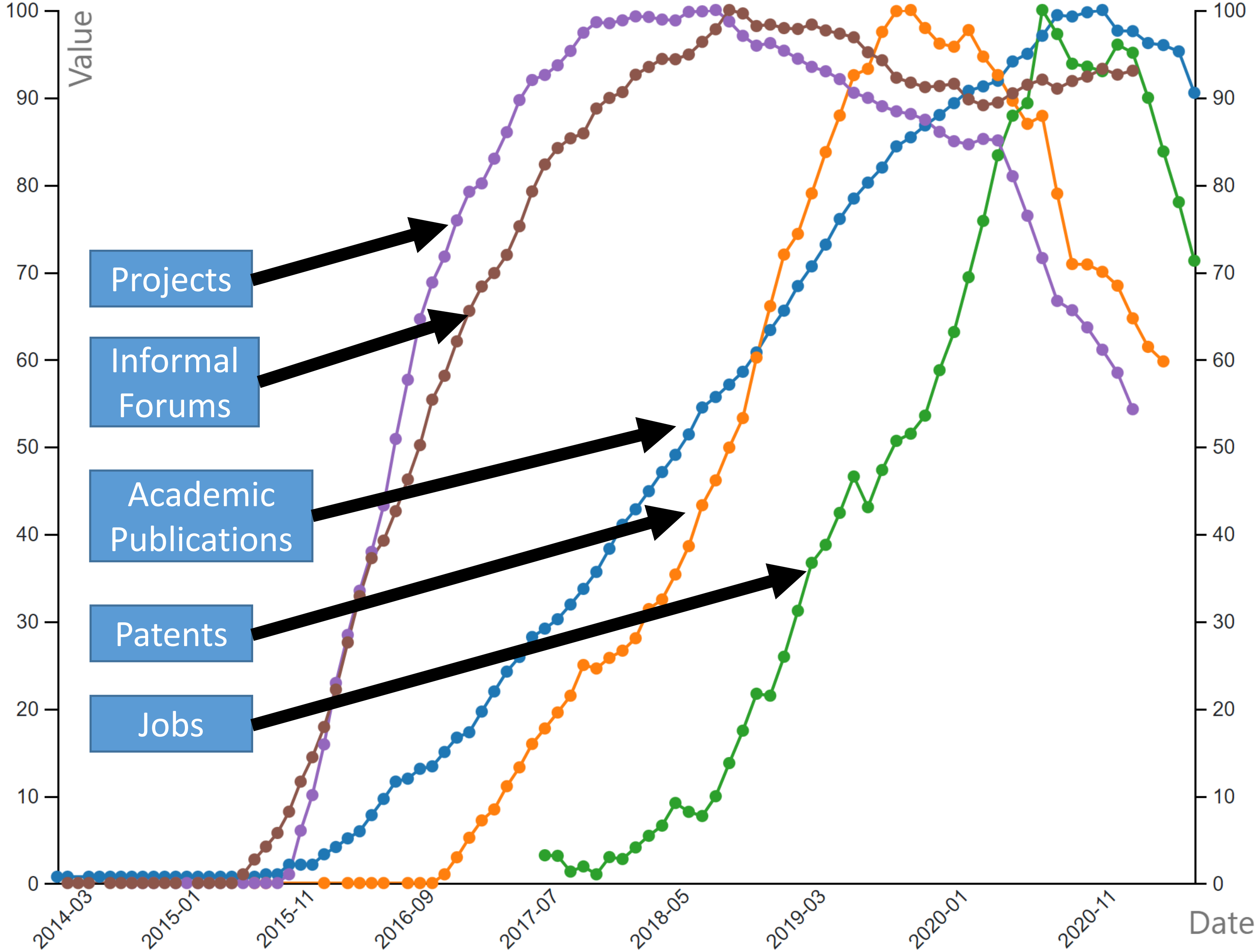


Academic Publications

Patents

Jobs

Cascading influence of an innovation (“LSTM algorithm” example)



Projects

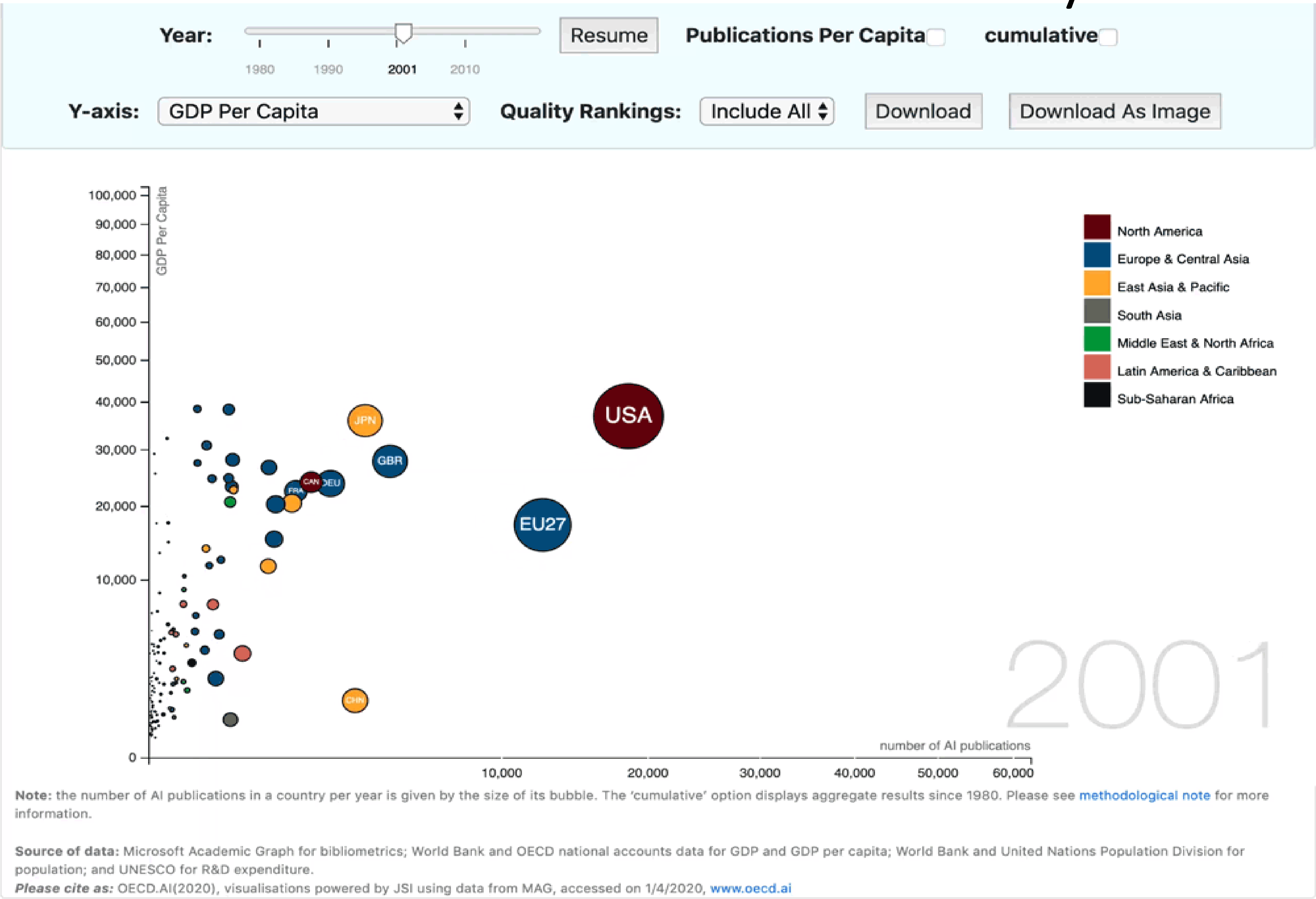
Informal Forums

Academic Publications

Patents

Jobs

Production of AI research over years



Year: 2001 Resume

Publications Per Capita □ cumulative □

Y-axis: GDP Per Capita

Quality Rankings: Include All

Download Download As Image

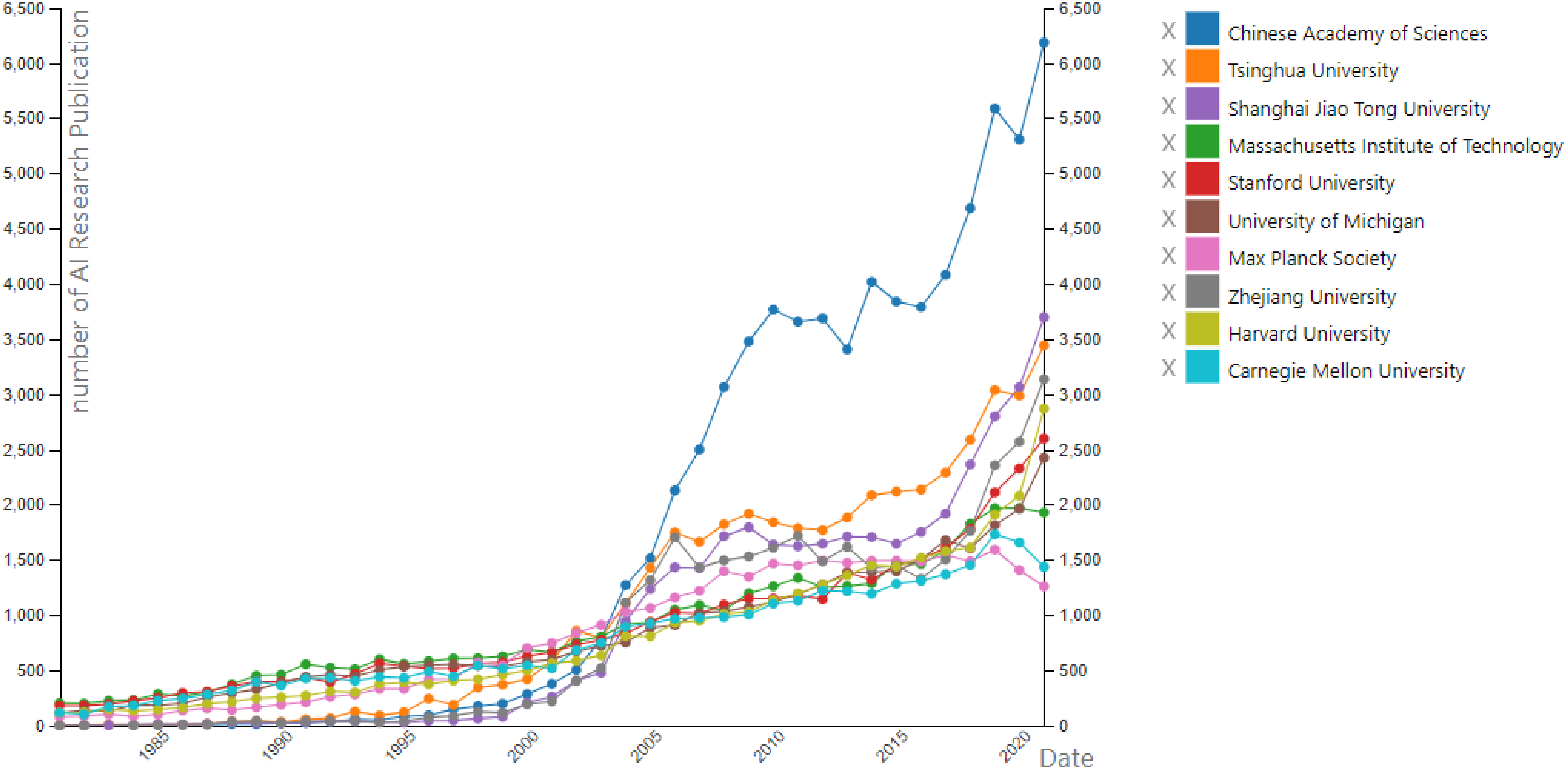
Note: the number of AI publications in a country per year is given by the size of its bubble. The 'cumulative' option displays aggregate results since 1980. Please see methodological note for more information.

Source of data: Microsoft Academic Graph for bibliometrics; World Bank and OECD national accounts data for GDP and GDP per capita; World Bank and United Nations Population Division for population; and UNESCO for R&D expenditure.

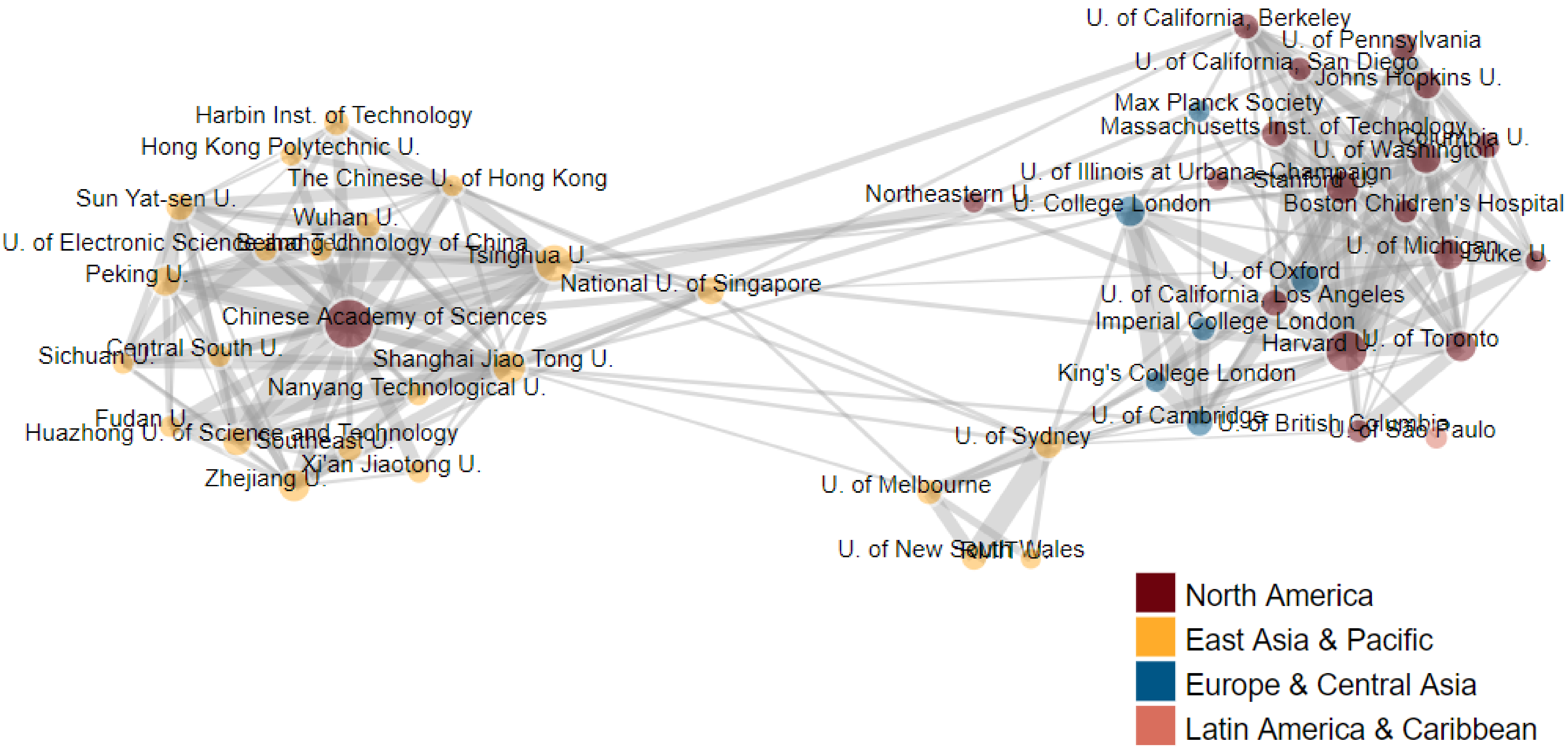
Please cite as: OECD.AI(2020), visualisations powered by JSI using data from MAG, accessed on 1/4/2020, www.oecd.ai

https://oecd-aiobservatory.ijs.si/dashboards/AI/Index?visualization=Research---MAG-Countries--ai-publications-vs-gdp-per-capita-by-country%2C-region%2C-in-time

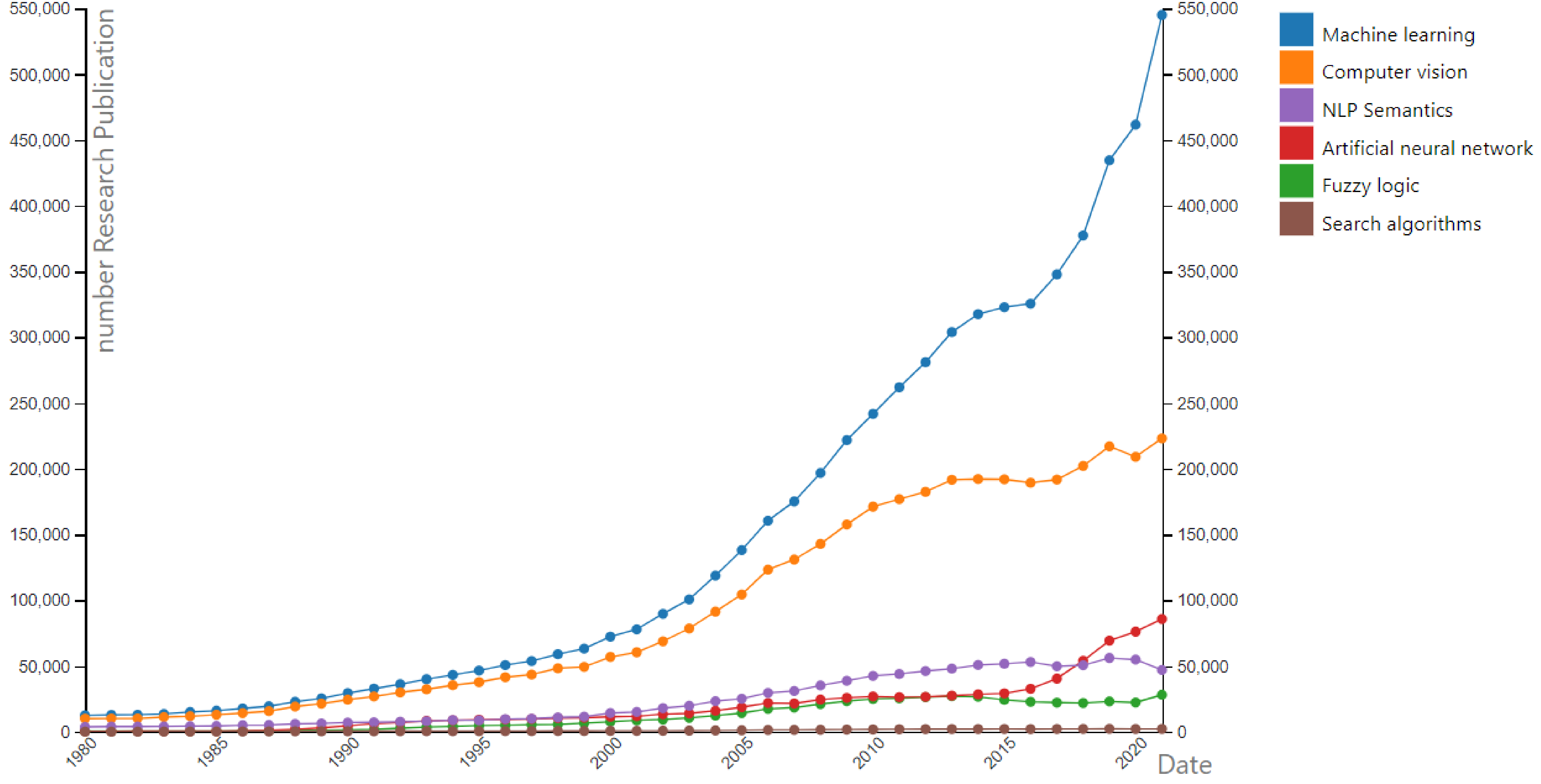
AI Research per institution



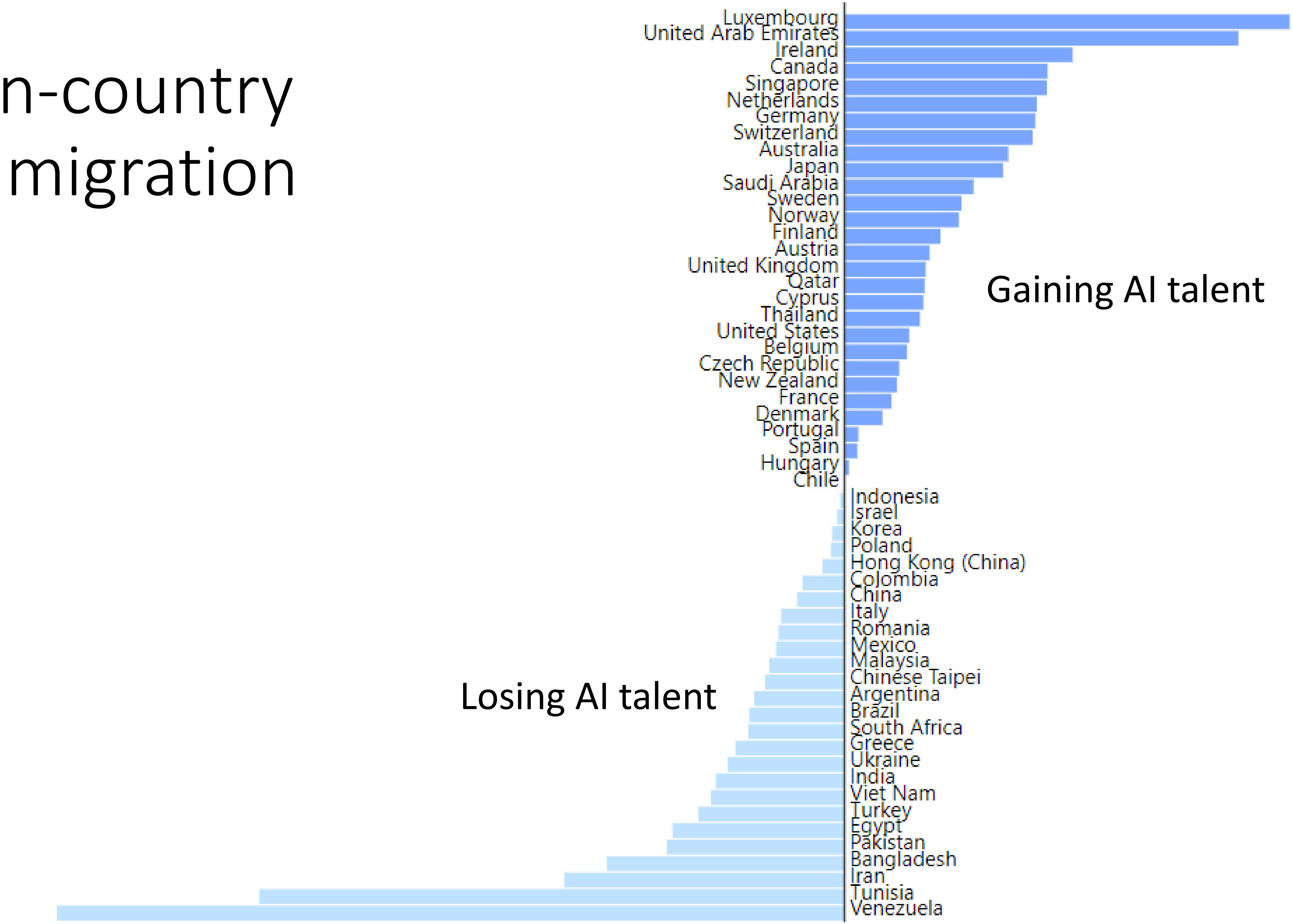
AI Research collaboration between institutions



Trends in AI subtopics over time



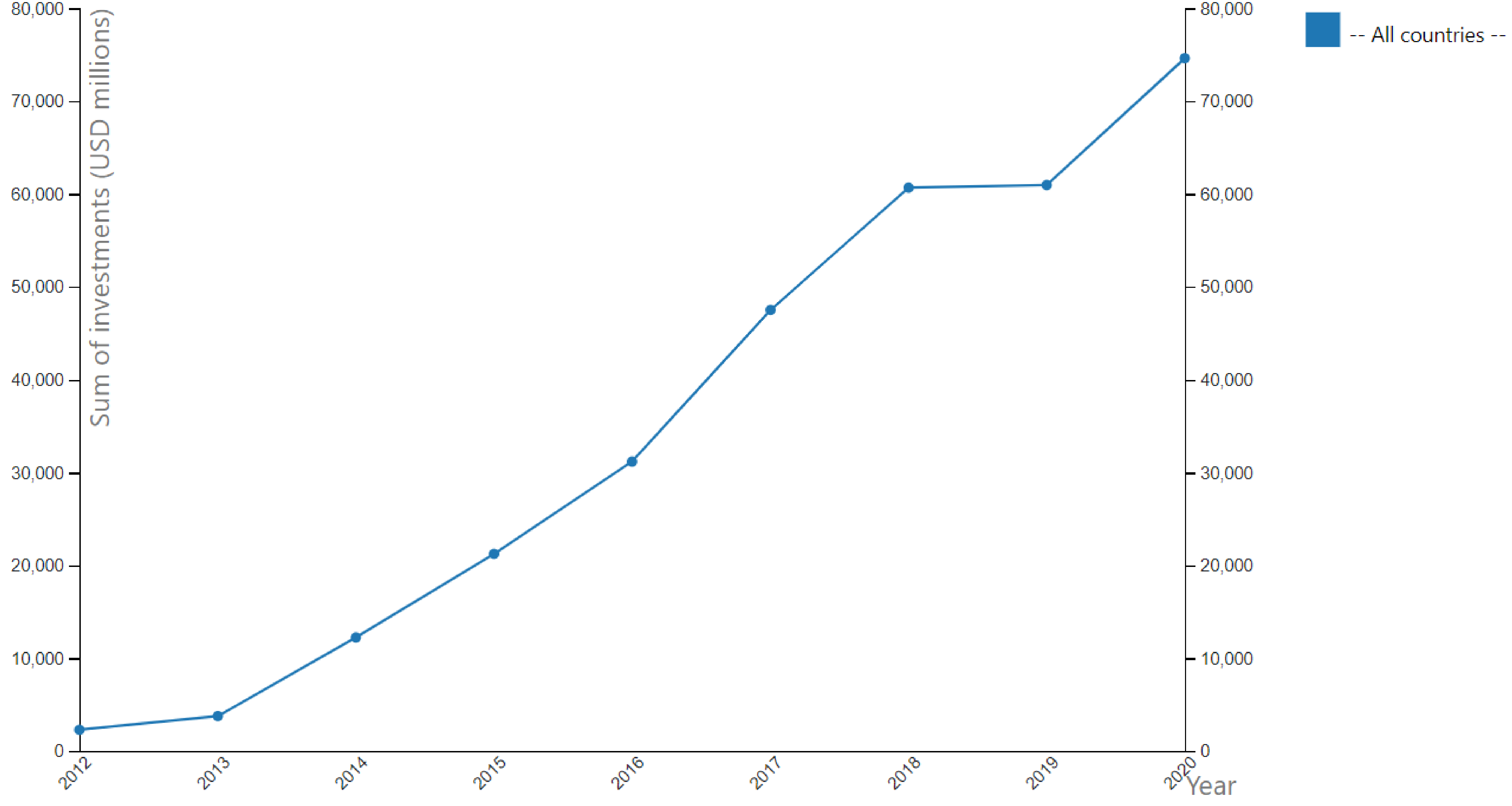
Between-country AI skills migration



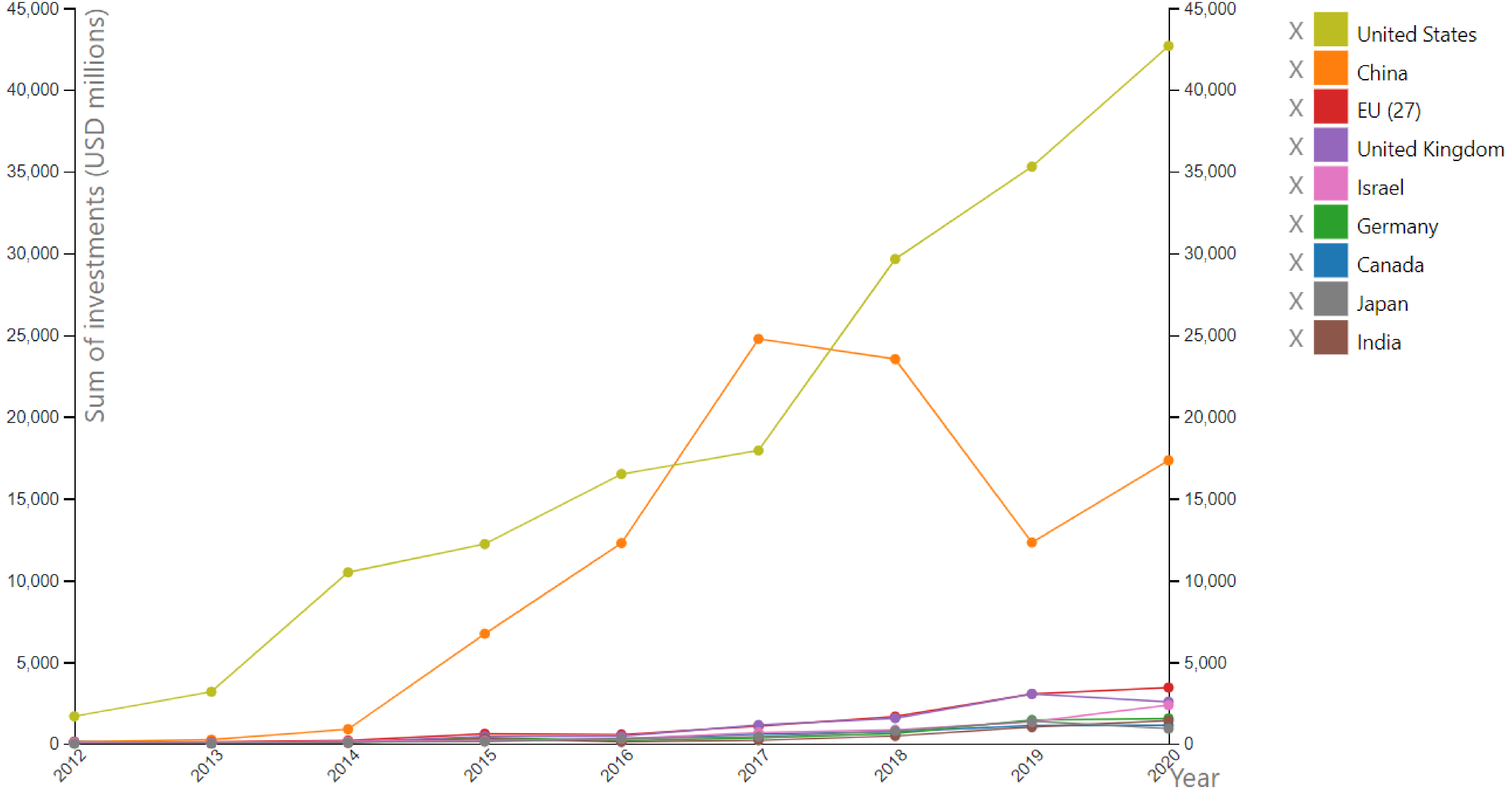
Top AI skills worldwide

1. Machine Learning
2. Artificial Intelligence (AI)
3. Data Structures
4. Deep Learning
5. NLP
6. Computer Vision
7. TensorFlow
8. Image Processing
9. Pandas
10. Scikit-Learn
11. Neural Networks
12. Keras
13. OpenCV
14. Artificial Neural Networks
15. PyTorch
16. Pattern Recognition
17. CNNs
18. Information Retrieval
19. Reinforcement Learning
20. Algorithm Development

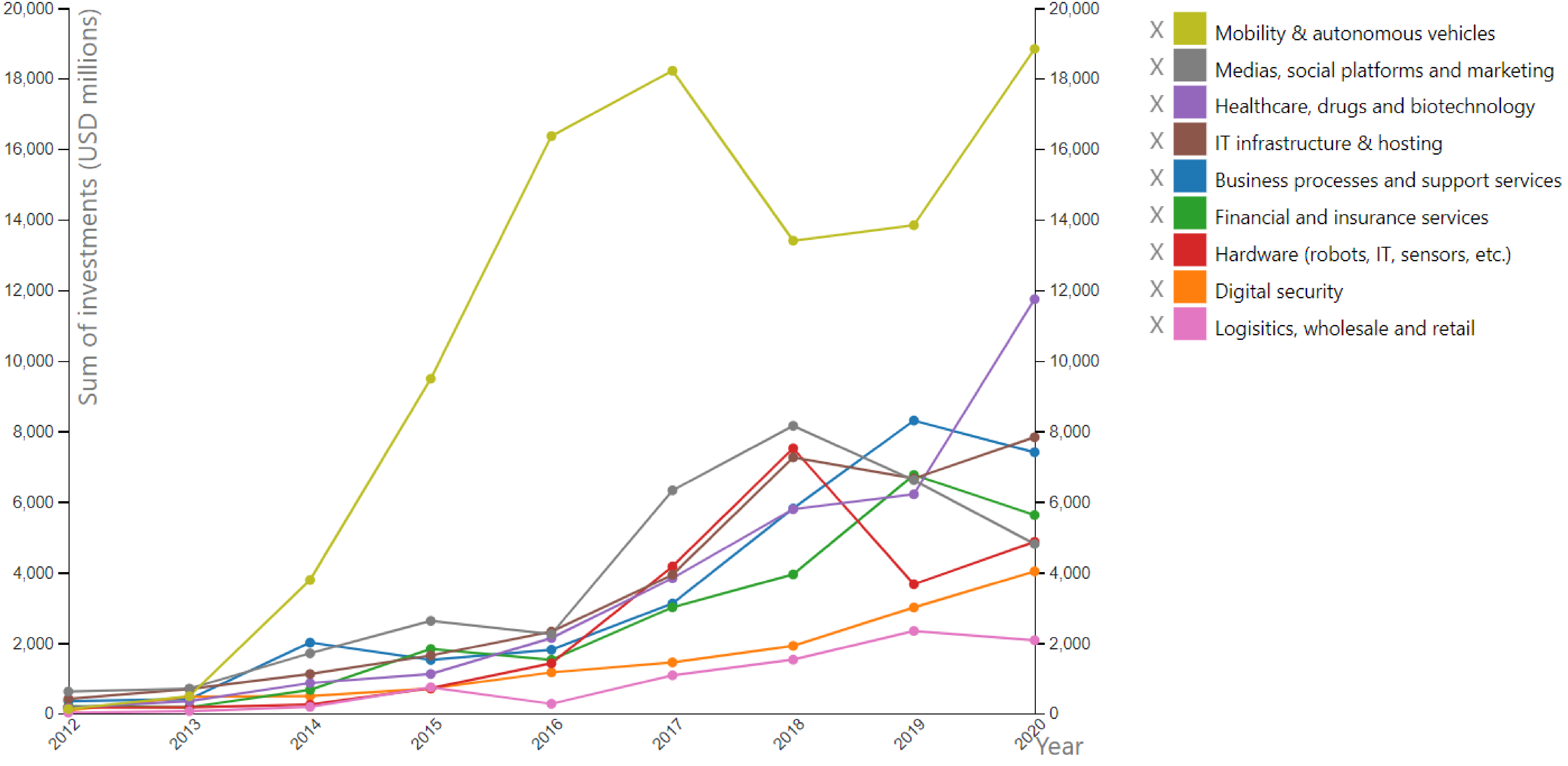
VC Investments in AI worldwide



VC Investments per country



VC Investments per AI sector



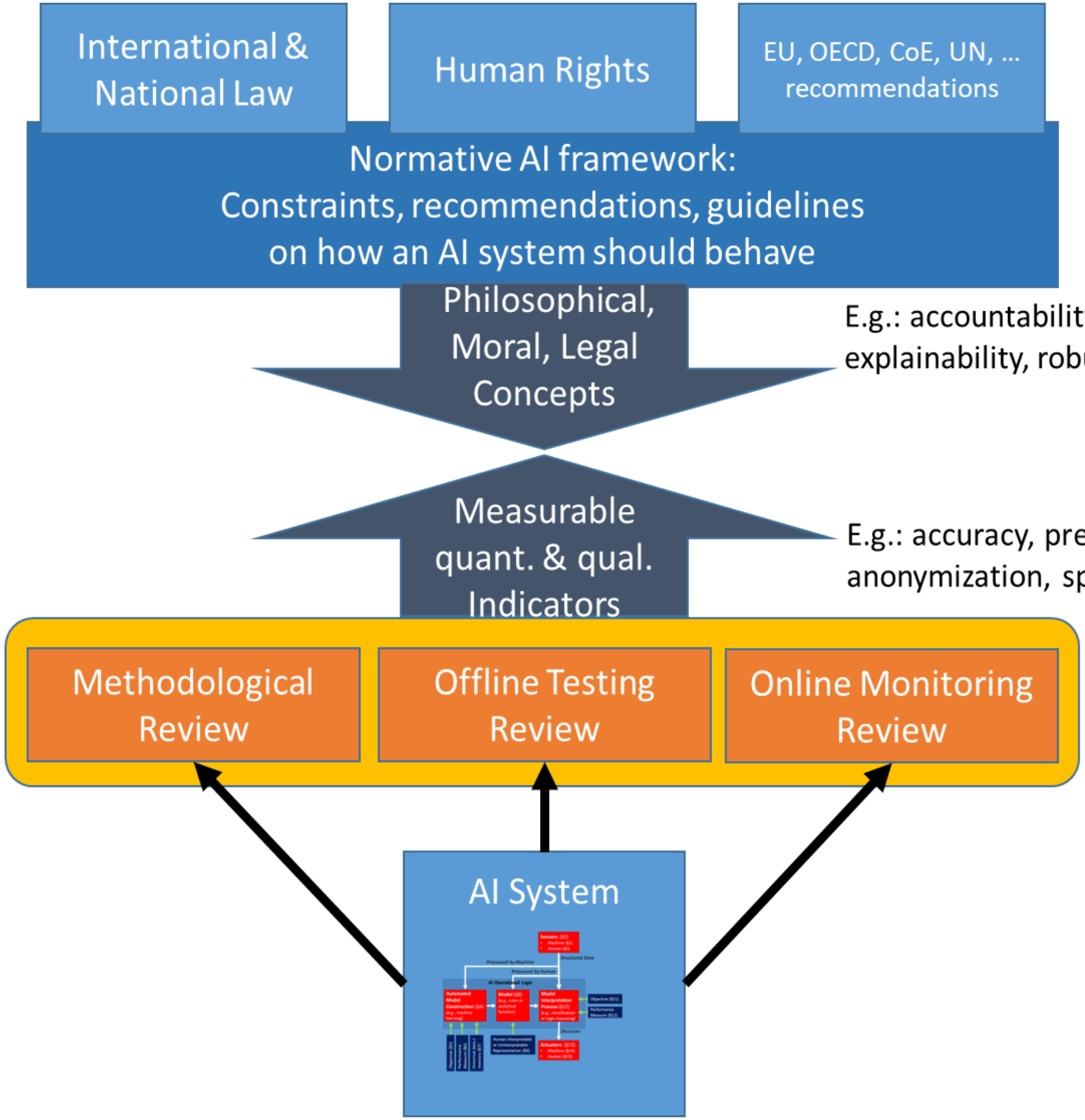
Some of the lessons learned from OECD AI Policy making

…semantic gap between legal & technical fields

…technology evolves faster as policy makers manage to regulate it

Normative vs technical indicators

- *High level view to the methodological approach on bridging the gap between normative systems (on the top) and technical AI systems (on the bottom).*
- *The gap appears between the abstract concepts used in normative documents and technical indicators measurable from a technical system.*



International & National Law

Human Rights

EU, OECD, CoE, UN, … recommendations

Normative AI framework: Constraints, recommendations, guidelines on how an AI system should behave

Philosophical, Moral, Legal Concepts

Measurable quant. & qual. Indicators

Methodological Review

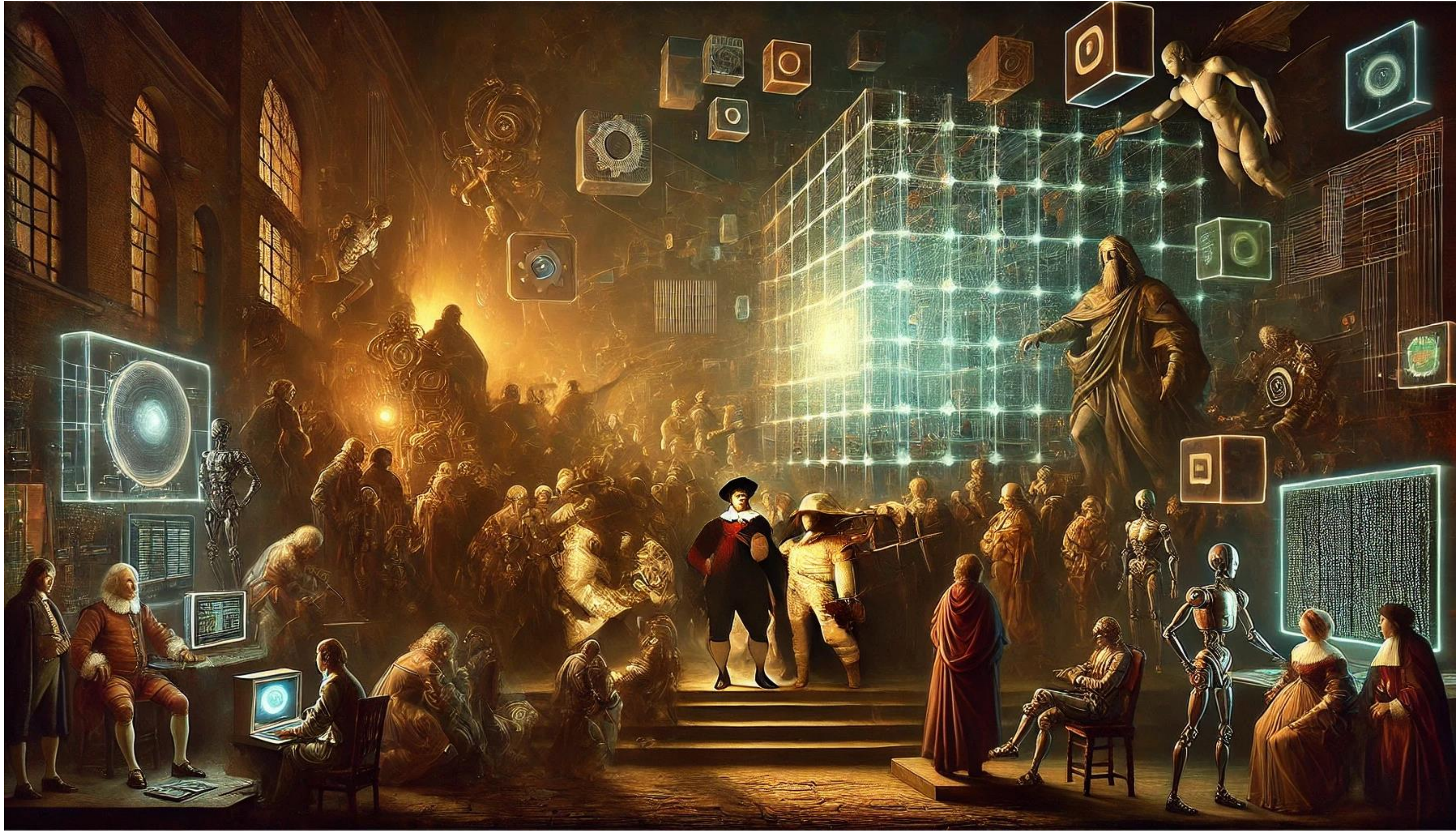
Offline Testing Review

Online Monitoring Review

AI System

E.g.: accountability, human centricity, fairness, transparency, explainability, robustness, security, safety, criticality, …

E.g.: accuracy, precision/recall, ROC/AUC, stat. significance, anonymization, speed, storage, application, industry sector, …



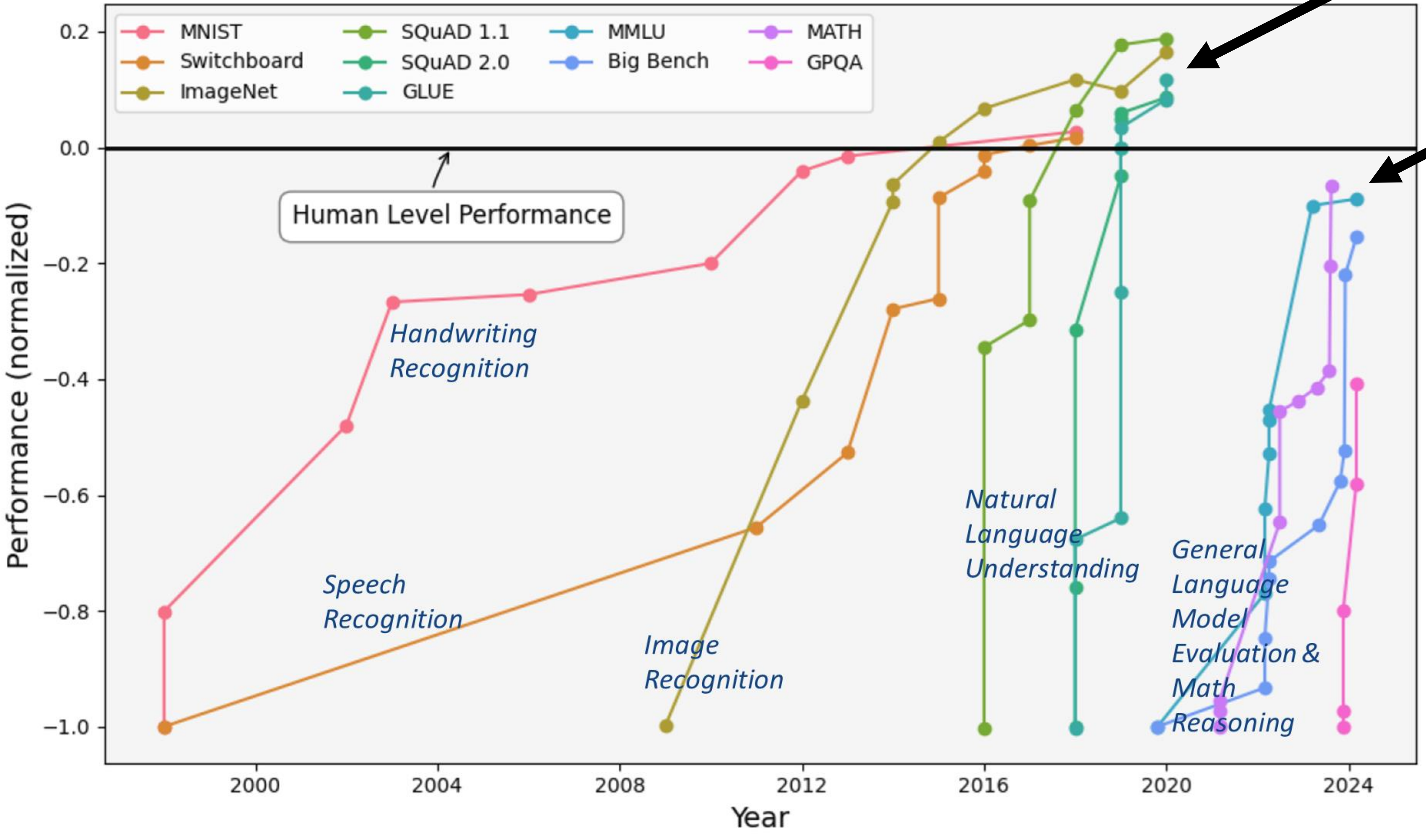
Technical triggers of AI & human rights

AI Systems vs. Human Capabilities

(Evolution of AI systems related to human skills)

Where we are
(basic human skills)

Where we go
(advanced
human skills)



https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

(Some of) the basic properties of AI systems which could endanger human rights

- **Managing large scale of complexity (recursive AI agents)**
  - …using the scale of data in the size of all human digital content
  - Humans cannot manage complexity beyond certain scale
- **Black-box models / lack of transparency**
  - …suitable for machine, but not for human
  - Humans don’t have feedback into machine (by explanation)
- **Speed of inferencing**
  - Surpassing humans in reaction time
  - The Speed of computers increases ~4 times per year
- **Autonomous Decision-Making (Human ‘Out of Loop’)**
  - …due to misalignment of human vs. machine value systems
- **Unclear accountability**
  - …the chain of stakeholders in the process is long
- **Robustness**
  - …AI systems are not perfect and is hard to guarantee stable results

(Un)Known-(Un)Knowns – Model Representation vs. Phenomena Discovery

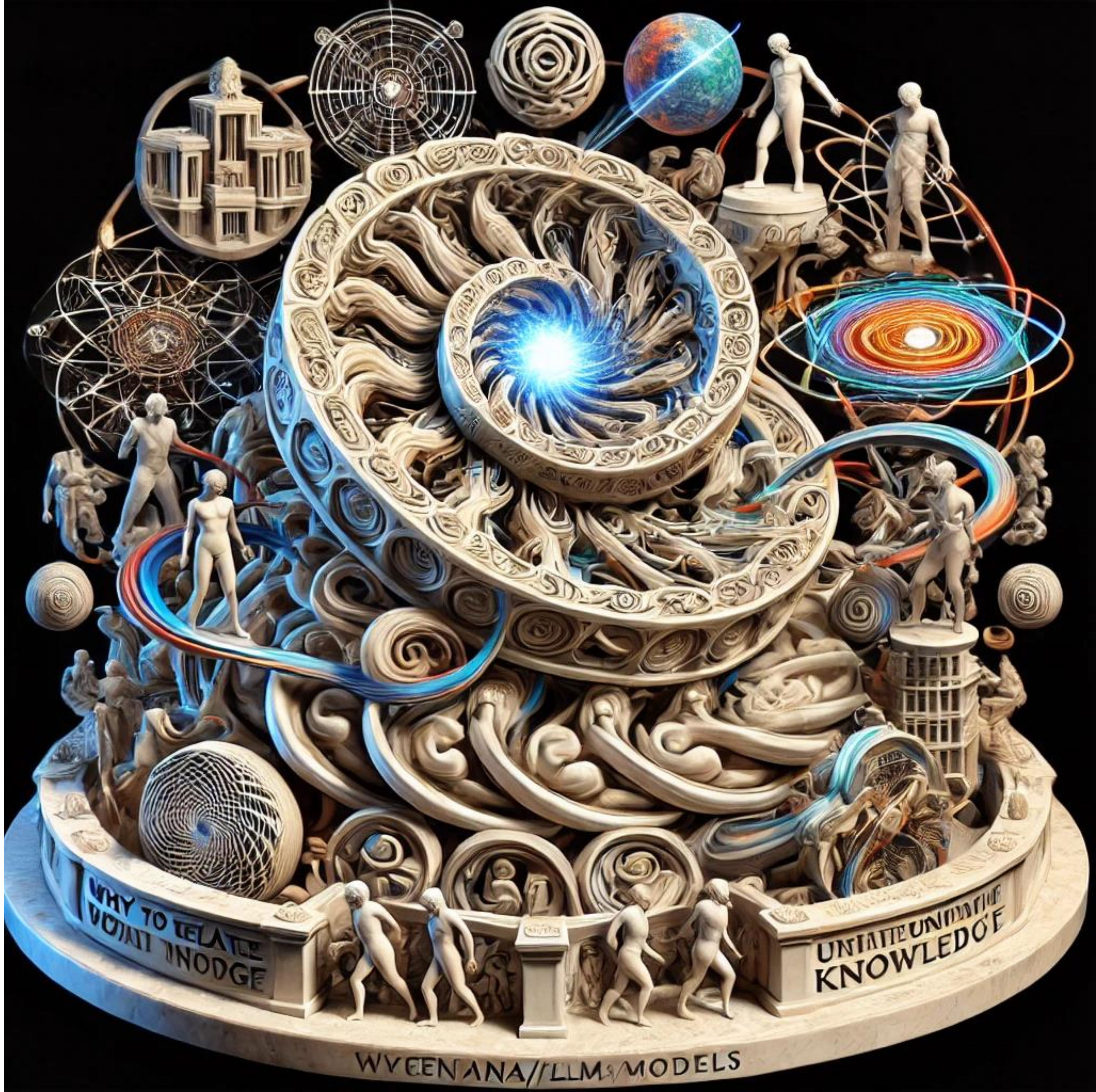
Phenomena Discovery

Model Representation

	Phenomena *Known* **to Humans** *(what people already know, but want to model and understand)*	**Phenomena** *Unknown* **to Humans** *(what people typically don’t know yet)*
Human Interpretable *(provided by a human to a machine)*	Traditional Statistics, Traditional AI, Logic Reasoning	Advanced Statistical Methods, Unsupervised AI (e.g. anomaly detection)
Human Uninterpretable *(created by a machine to optimize the solution)*	Modern AI (after 2010), Deep Neural Networks, Transformers, Reinforcement Learning	AI to come, e.g., AI with “multihop” reasoning, Online Reinforcement Learning

…this would allow to reach yet undiscovered concepts and relations and reach insights far from what humanity knows today

Likely future AI development



Speed of computers: computers are ~4 times faster every year

- If computers will be expectedly much faster in the near future, what can we do with such capacity?
- …what fundamental AI problems could be addressed and what consequences this could have?

Jensen Huang, NVIDIA CEO, March 19, 2024:

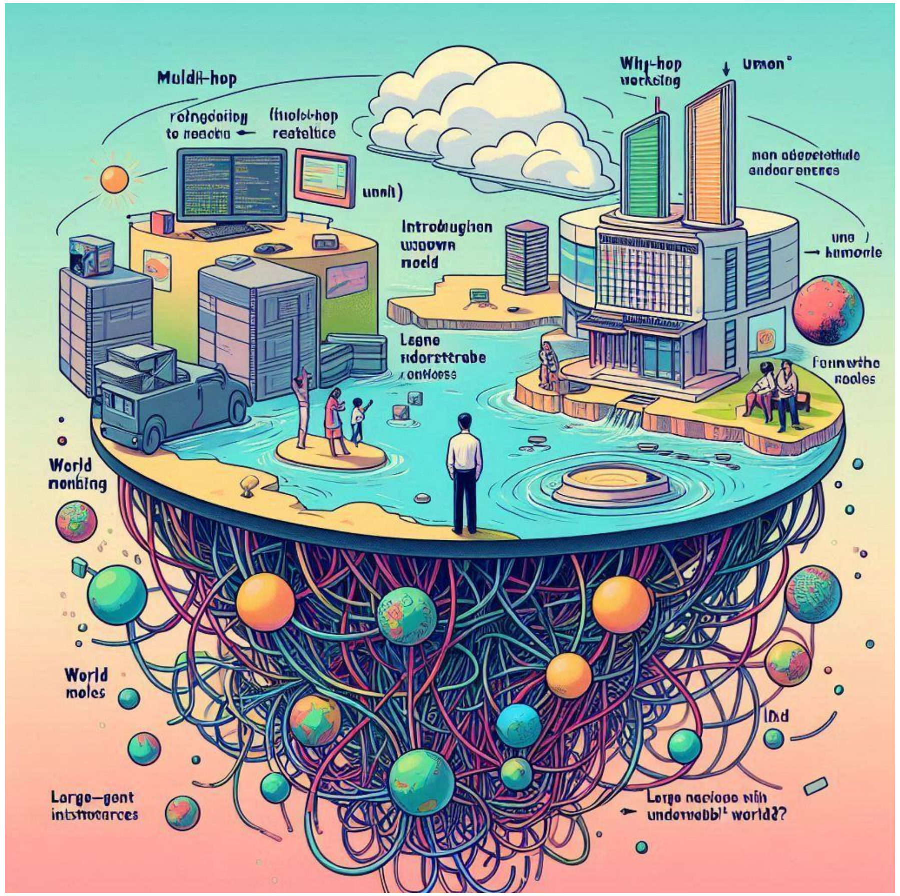
"Moore's Law, in its best days, would have delivered 100x in a decade," Huang explained. "By coming up with new processors, new systems, new interconnects, new frameworks and algorithms and working with data scientists, AI researchers on new models, across that entire span, we've made large language model processing a million times faster."

https://siepr.stanford.edu/news/nvidias-jensen-huang-incredible-future-ai

https://www.youtube.com/watch?v=cEg8cOx7UZk

(Near- to Mid-)future AI challenges

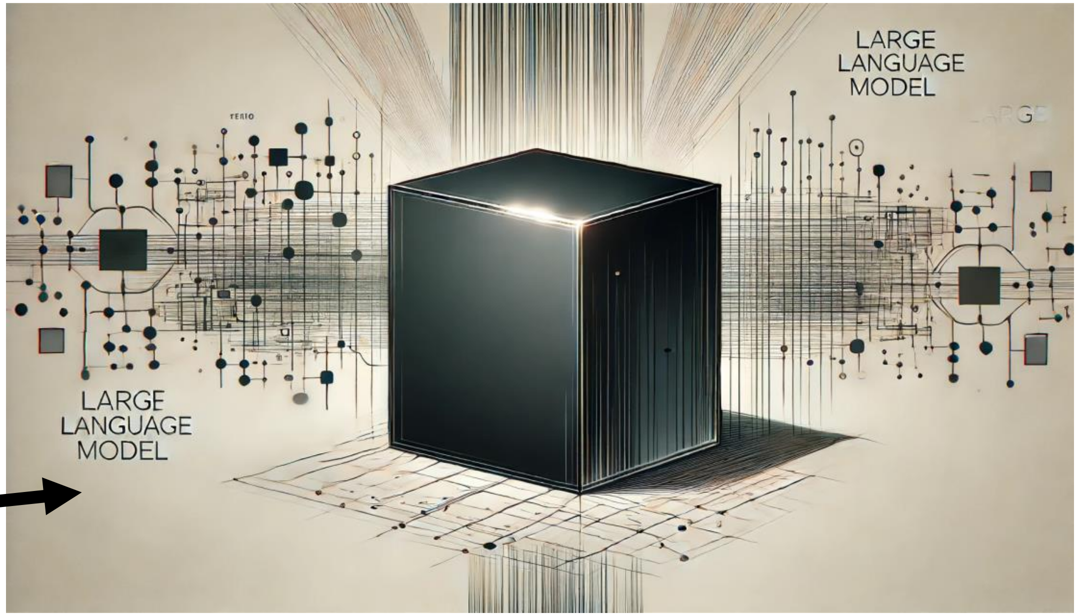
1. **Advanced reasoning capabilities** to reach *(un)known (un)known knowledge*
2. **Why** GenAI/LLMs works at all?
3. Introducing “**World Models**“ to relate with human understandable world
4. Large **(recursive) AI agent** infrastructures with autonomous emergent behaviors
5. Integrating new **data modalities (types of data)** beyond the usual ones



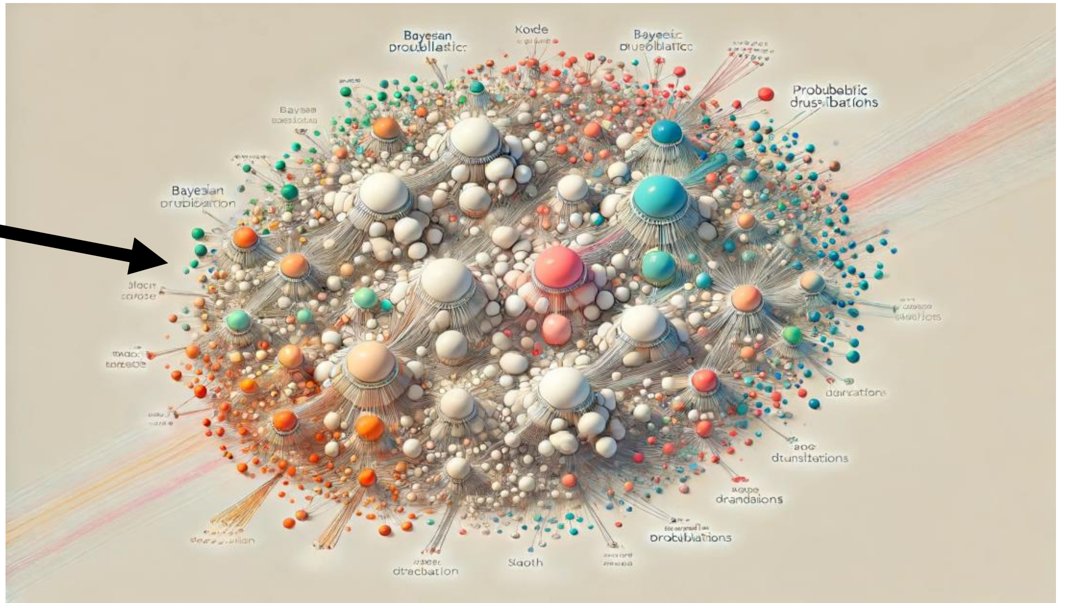
How LLM models see the world?

- There is no explicit “***world model***”
  - …i.e., machine does not understand the world

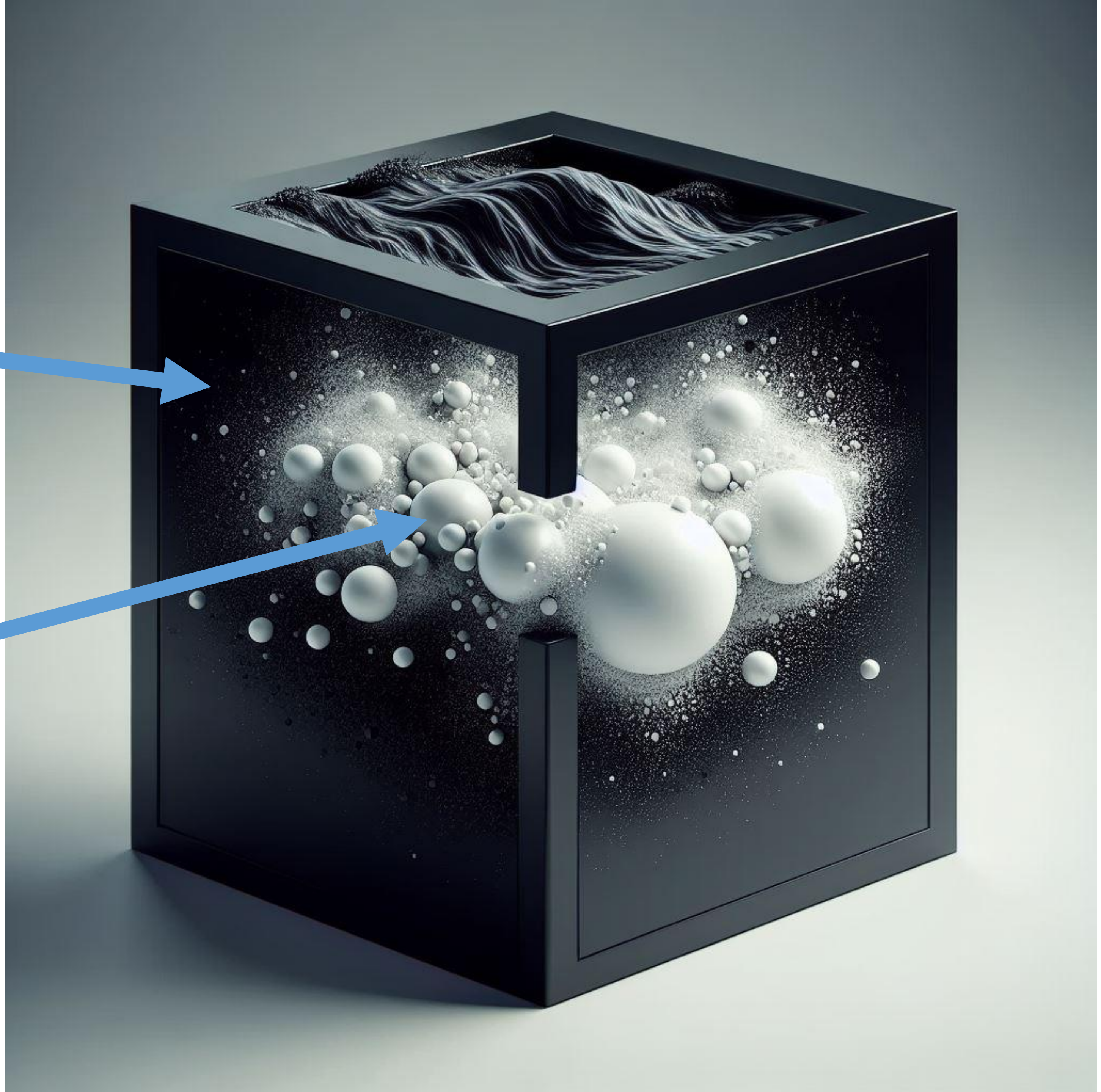
- For humans it looks like a “***big black-box***”
  - …since it is expressed in a language not understandable by humans



- Internally the black box is a huge **network of interleaved probabilistic concepts**
  - …could be visualized as a network of interconnected clouds representing concepts



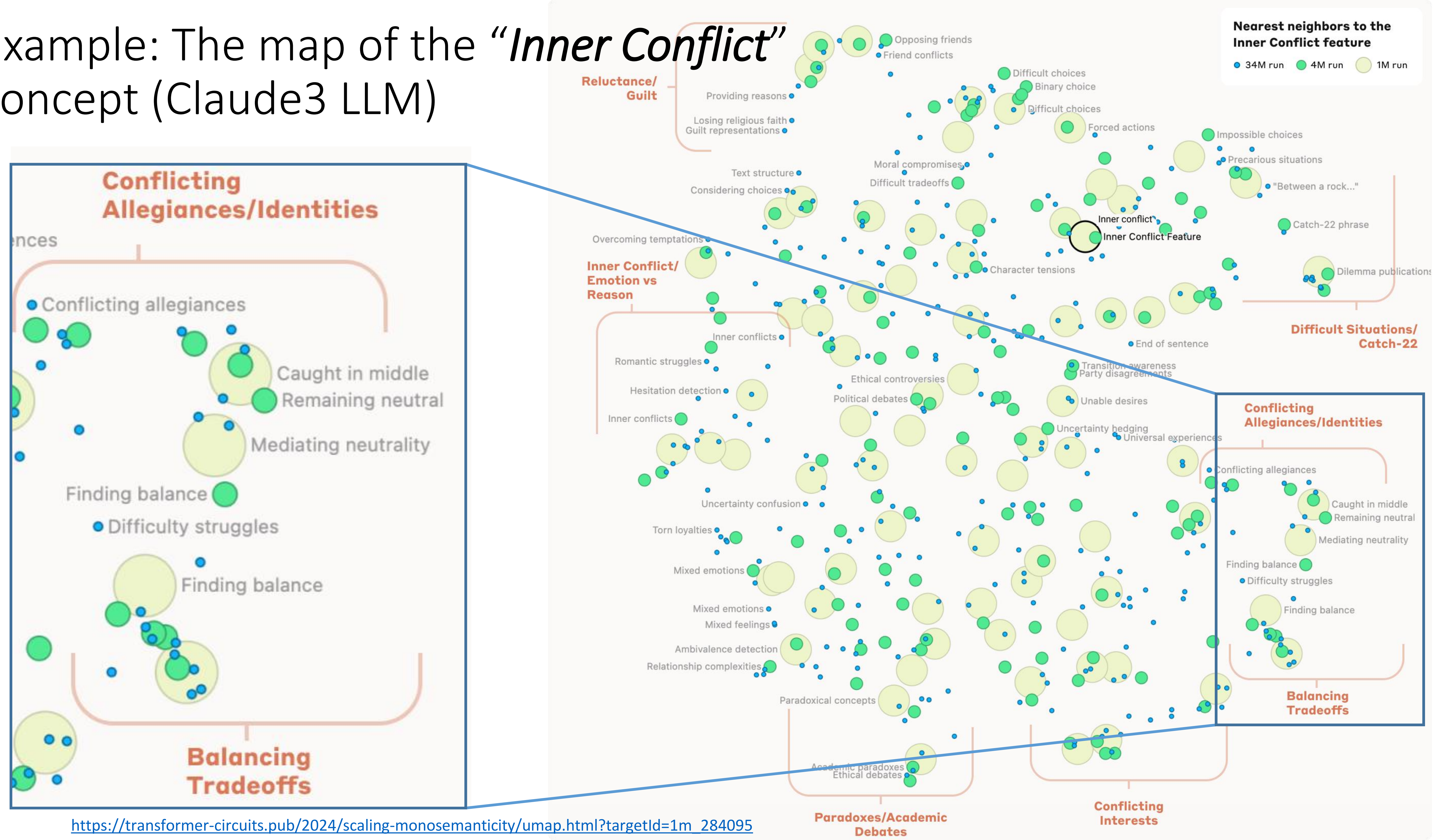
LLMs & World Models



LLM as a big black-box

Fragments of explainable knowledge
(via local “world models”)

Example: The map of the "*Inner Conflict*" concept (Claude3 LLM)



https://transformer-circuits.pub/2024/scaling-monosemanticity/umap.html?targetId=1m_284095

Questions?