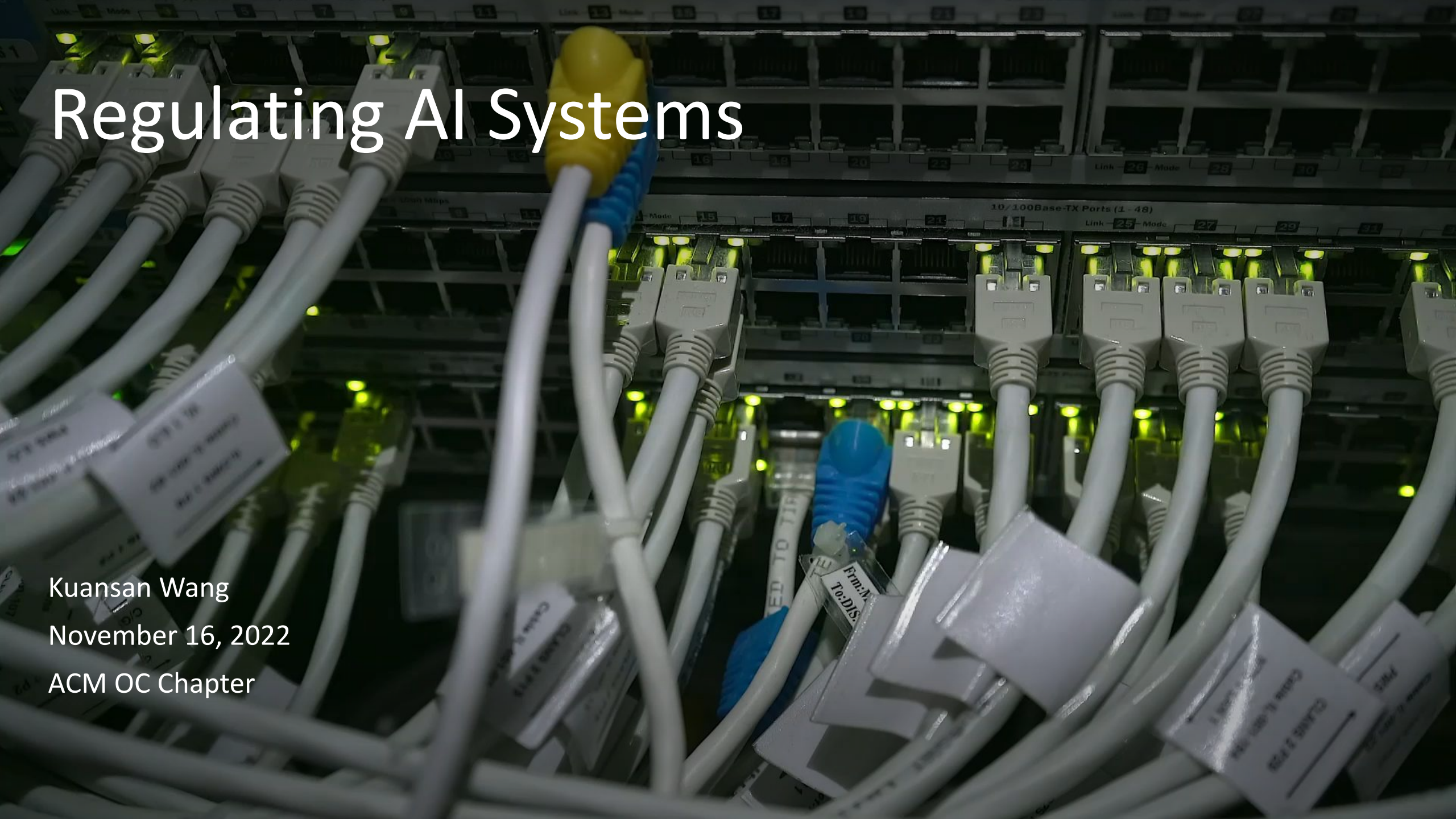
Regulating AI Systems

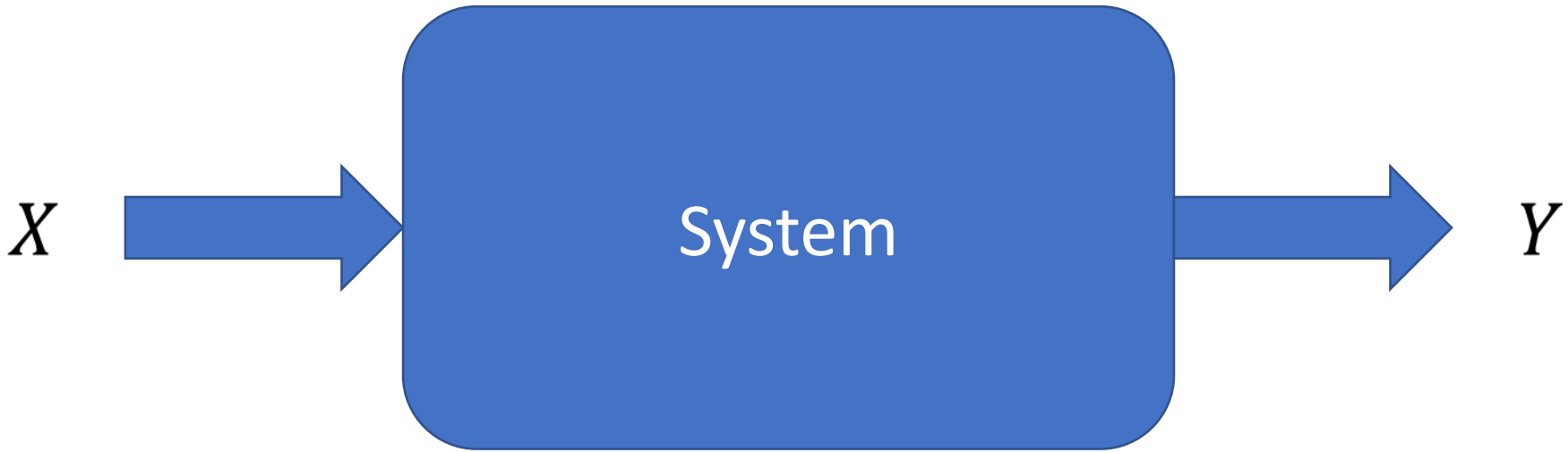


Kuansan Wang

November 16, 2022

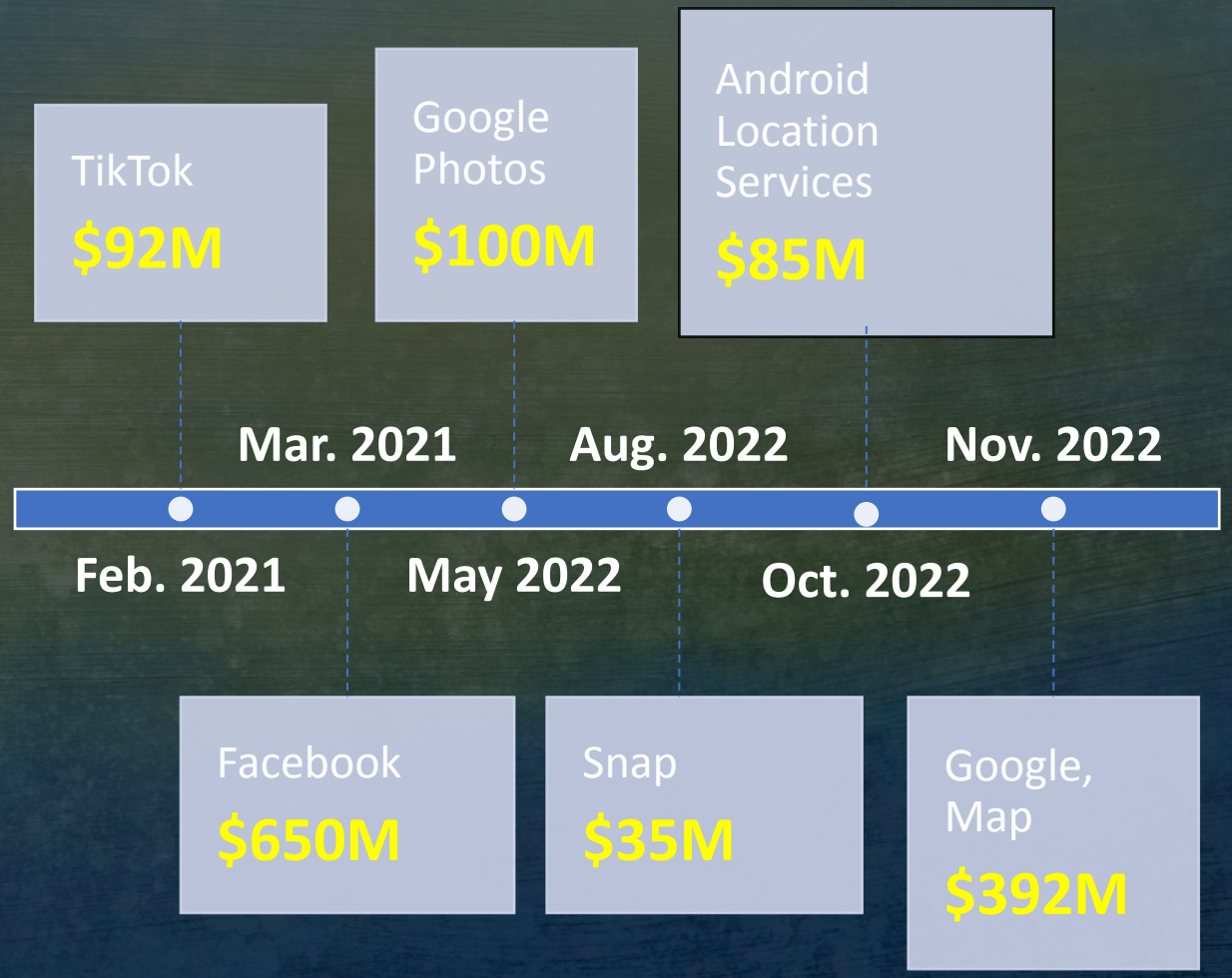
ACM OC Chapter

Decision Theory View on AI Systems



$$\arg\max_{Y \in \Omega} E[U(X,Y)P(Y|X)]$$
$$= \arg\max_{Y \in \Omega} E[U(X,Y)P(X|Y)P(Y)]$$

Recent US Lawsuits and Fallouts



THE WALL STREET JOURNAL.

Tech Giants' New Appeal to Governments: Please Regulate Us

Facing probes and growing public backlash, top leaders at Microsoft, Apple, Google and Facebook call publicly for new laws

By Sebastian Herrera

Jan. 27, 2020 7:01 am ET

EU's AI Act(AIA) Proposal

- Risk assessment & monitoring on AI system with predefined obligations
- Unacceptable risk: "AI systems with clear threat… will be banned"
  - Manipulating human behaviors
  - Circumvent users' free wills
- Examples of "high risk" systems:
  - Transportation
  - Education (e.g., exam scoring)
  - Employment (e.g., CV screening)
  - Essential services (e.g., financial, healthcare)
  - Law enforcement and judiciary

EUROPEAN COMMISSION

Brussels,
21.4.2021

COM(2021)
206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

Responses to EU AIA

Google

- Definition of high risk
- Balance between ex ante assessments and ongoing due-diligent reviews
- Principles and process

Microsoft

- Impacts rather than scenarios
- Calibrate obligations of different actors
- Principles and process

WIRED SEP 1, 2021

The Fight to Define When AI Is 'High Risk'

Everyone from tech companies to churches wants a say in how the EU regulates AI that could harm people.

Other governmental activities

US OSTP, Blueprint for an AI Bill of Rights (Oct 2022)

UNESCO, Recommendation on the Ethics of AI (Nov 2021)

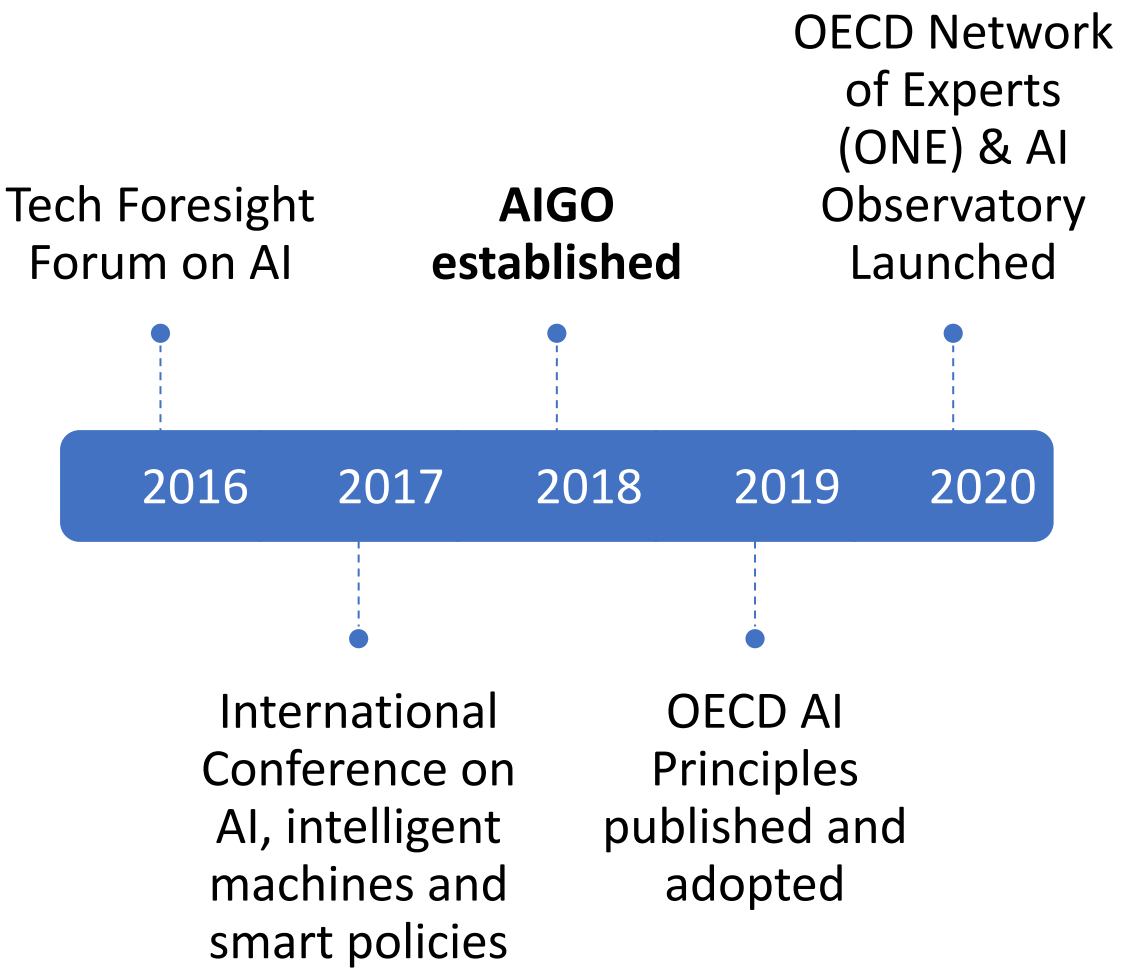
G20, AI Guidelines (June 2019)

OECD, AI Principles (May 2019)

OECD AI Principles

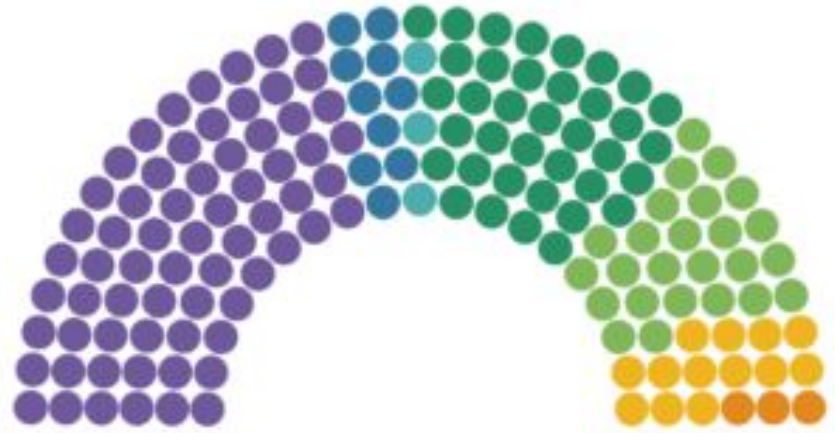
- Inclusive growth, sustainable development
- Human centered values, fairness
- Transparency, explainability
- Robust, secure, safe
- Accountability

OECD AI Governance (AIGO)



ONE Members

- Governments, International Orgs
- Businesses
- Civil Societies & Academia
- Technical Community
- Trade Unions

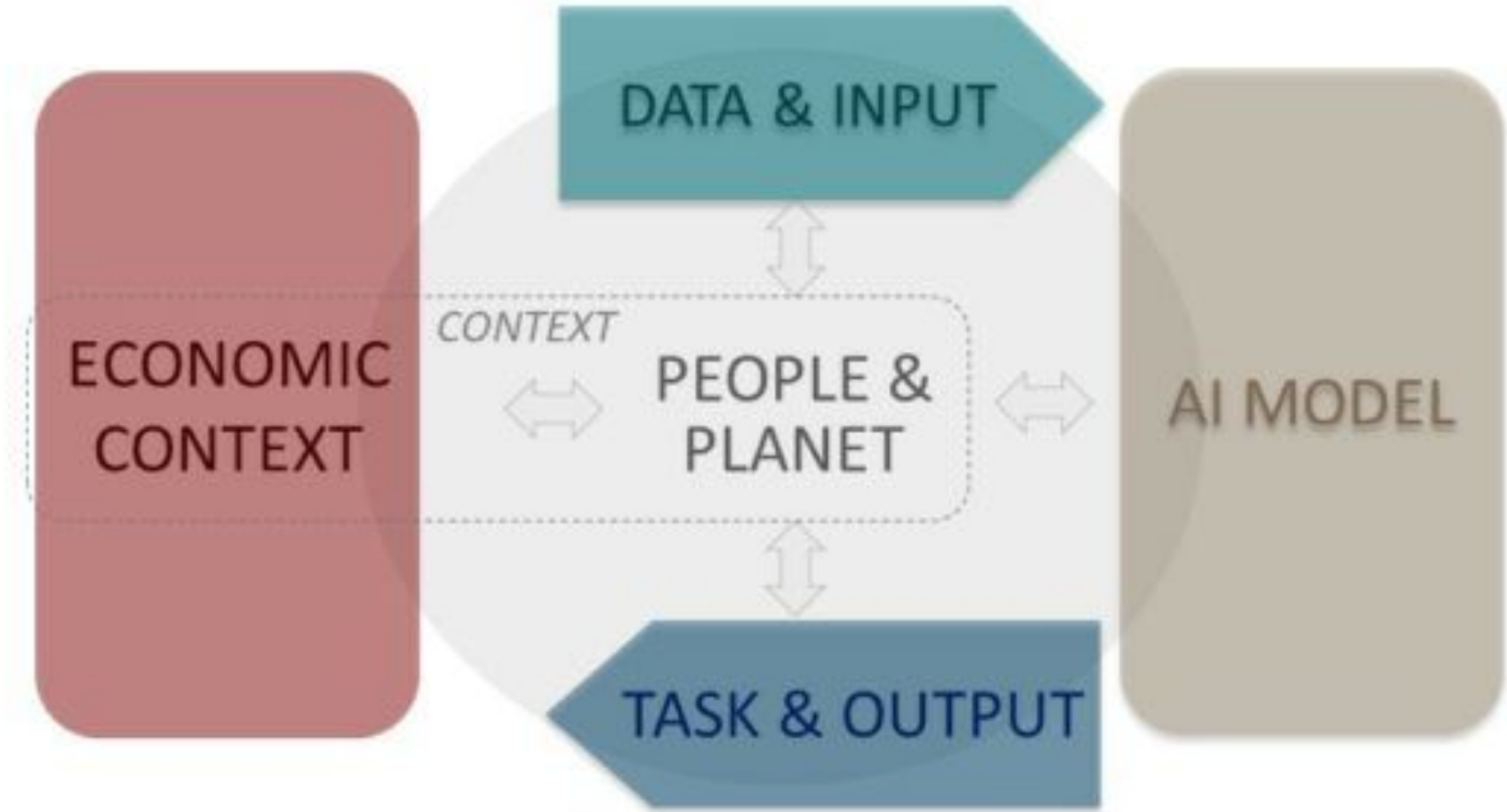


ONE Working Groups

Classification & Risk

Tools & Accountability

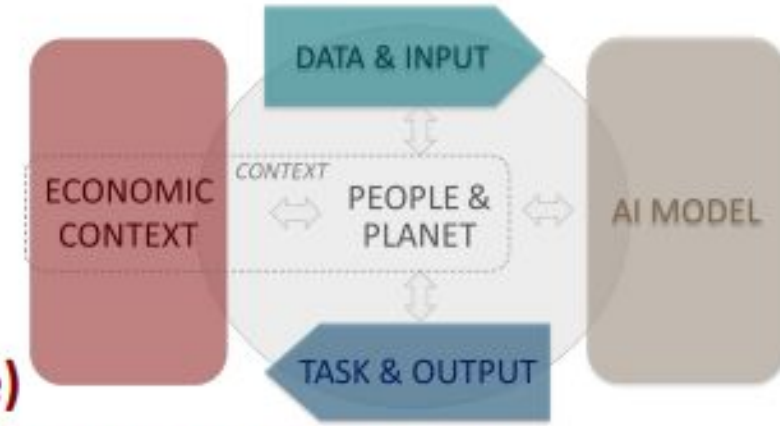
Compute & Climate


DATA & INPUT
ECONOMIC CONTEXT
CONTEXT
PEOPLE & PLANET
AI MODEL
TASK & OUTPUT

Example 1: Credit-scoring AI systems

OECD.AI Policy Observatory

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)

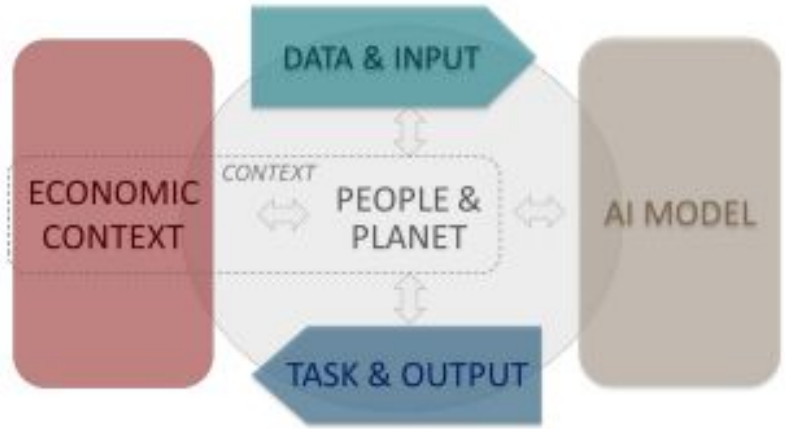
Example 2: GPT-3, text generation

OECD.AI Policy Observatory

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]



Next steps at the OECD:

OECD.AI Policy Observatory

- **Refine classification criteria**
  - Add more real-world AI systems and identify possible indicators
- **Develop a risk assessment framework to facilitate global interoperability**
  - leveraging the classification plus possible governance at the corporate, institution or AI systems level
  - Leveraging work in partner organisations, including EU, US, ISO
  - Leveraging risk assessment work in other parts of the OECD
  - Develop a common framework for reporting about AI incidents.
- **Support risk management**: Inform related work on mitigation, compliance and enforcement along the AI system lifecycle, and responsible business-impact assessment.

The OECD Artificial Intelligence Policy Observatory - OECD.AI