OECD.AI
Policy Observatory

AI System Classification for Policymakers

ONE AI WG on AI System Classification

with Dewey Murdick

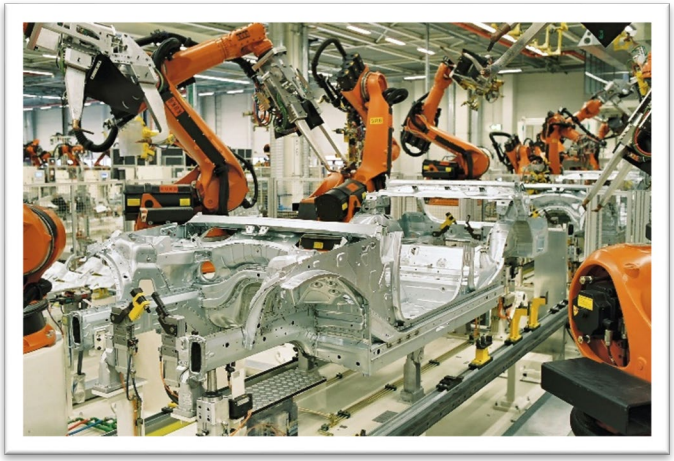
A platform to share and shape Artificial Intelligence policies

OECD
BETTER POLICIES FOR BETTER LIVES

OECD.AI Policy Observatory

Why classify AI systems?

Different types of **AI systems raise unique policy considerations** in their user context



OECD.AI Policy Observatory

Policy-Impacting Events

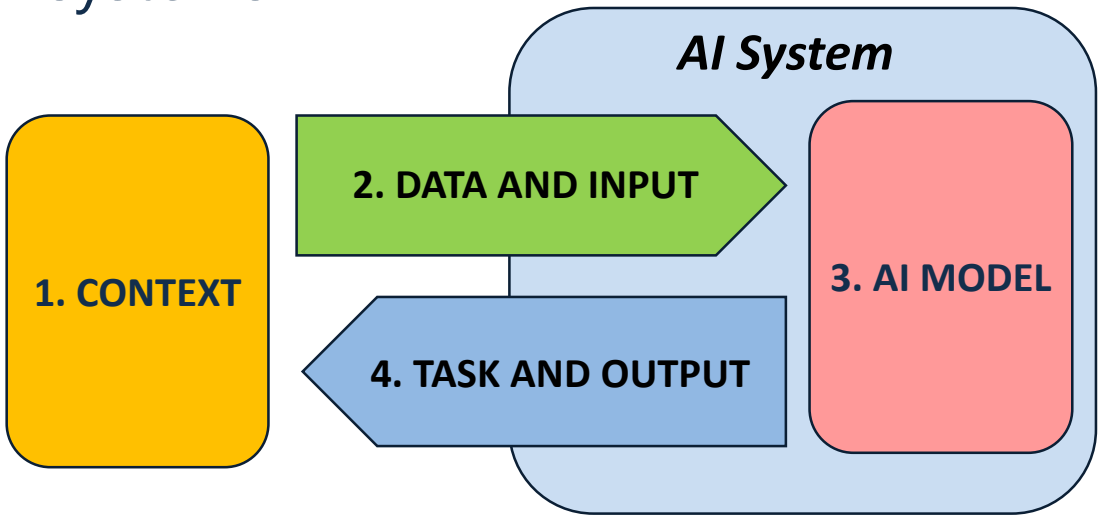
Consider AI Incidents – unforeseen failures of intelligent systems – for example

- An autonomous car kills a pedestrian
- A trading algorithm causes a market "flash crash" where billions of dollars transfer between parties
- A facial recognition system causes an innocent person to be arrested
- *… and the fundamental data gaps (e.g., women, minorities) that impact systems and*

AI systems cannot be treated as a single type of technology, policymakers must be contextually aware

Overview of the classification framework

A user-friendly framework to navigate policy implications of different *types* of AI systems



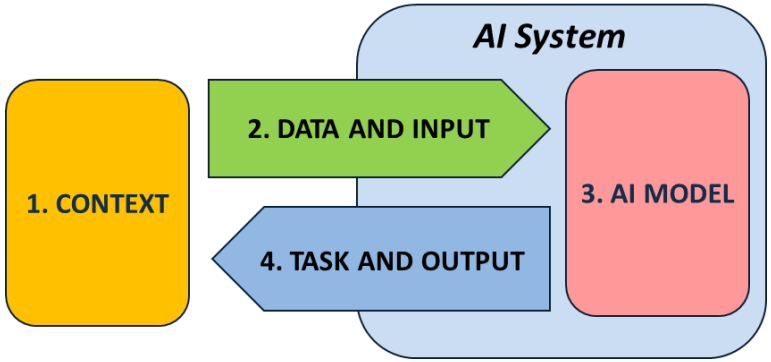
4 key dimensions:

1. **Context**, including sector (healthcare, etc.), impact and scale
2. **Data and input**, including data collection, personal nature of data
3. **AI model (technologies)**, incl. model type and model building process
4. **Task and output**, incl. AI system's task (e.g., recognition, personalisation, etc.) and action autonomy

OECD.AI Policy Observatory

Key Elements to Consider (selected)

- **Sector of deployment**
  (e.g., Transportation and storage, Human health and social work activities, Education)
- **Critical function**
  (e.g., health, safety, and security of citizens; essential economic and societal services)
- **System users**
  (e.g., AI-expert vs. non-AI expert user)
- **Data collection**
  (e.g., humans, automated, system experience)
- **Data domain**
  (e.g., proprietary, public, personal)

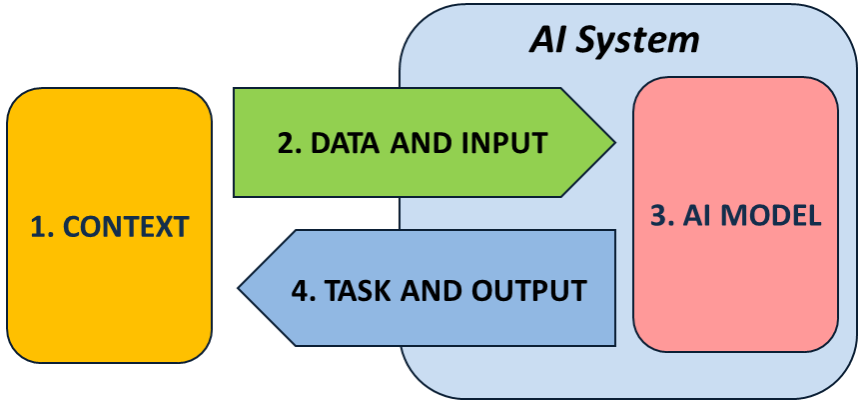


- **Acquisition of capabilities**
  (e.g., learn from people vs. provided data vs. system experience)
- **System task**
  (e.g., recognition, personalization, goal-driven optimization)
- **Level of action autonomy**
  (e.g., high (human out-of-the-loop), medium (human on-the-loop), low (human-in-the-loop)

OECD.AI Policy Observatory

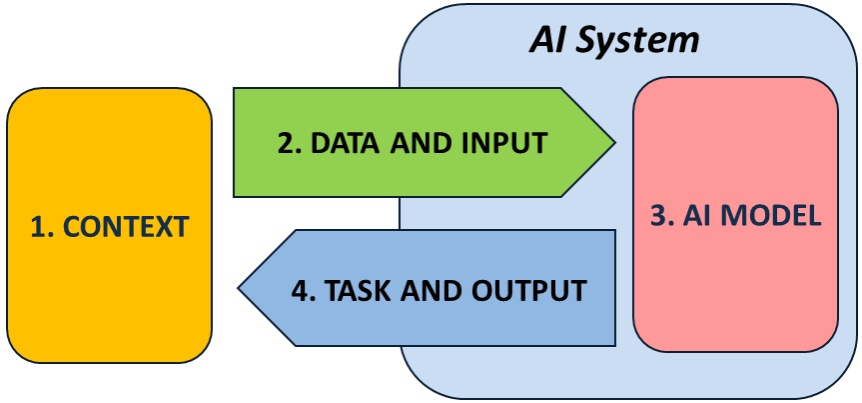
AlphaGo Zero, games (example)

- **Sector of deployment**
  Arts, entertainment and recreation
- **Critical function**
  None
- **System users**
  Original users are AI expert user
- **Data collection**
  Automated source
- **Data domain**
  Proprietary data source
- **Acquisition of capabilities**
  Learn from system experience
- **System task**
  Goal-driven optimization – giving systems a goal and the ability to find the optimal solution
- **Level of action autonomy**
  High (human out-of-the-loop)



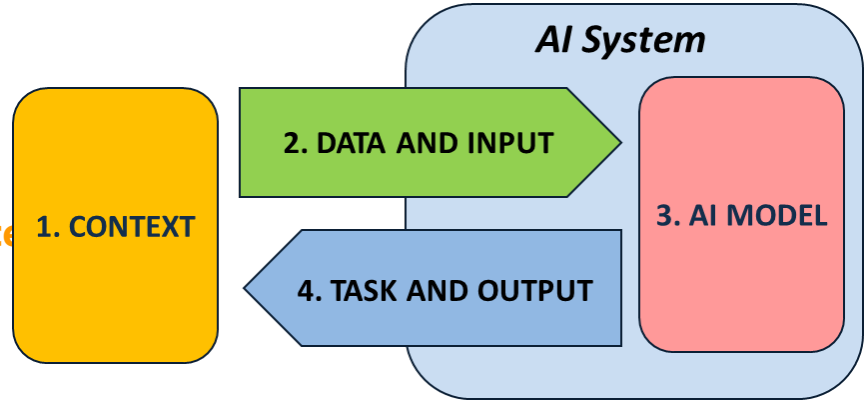
GPT-3, text generation (example)

- **Sector of deployment**
  Information and communication
- **Critical function**
  None
- **System users**
  Primary users are non-AI expert user
- **Data collection**
  Human sources
- **Data domain**
  Public data sources
- **Acquisition of capabilities**
  Learn from provided data
- **System task**
  Goal-driven optimization – giving systems a goal and the ability to find an optimal solution
  Interaction support – creating content to power machine-human interaction
- **Level of action autonomy**
  Medium (human on-the-loop) [human action required (e.g., use of generated text)]



Credit-scoring AI systems (example)

- **Sector of deployment**
  Financial and insurance activities
- **Critical function**
  Critical function/activity (economic service
- **System users**
  Primary users are non-AI expert user
- **Data collection**
  Human and automated sources
- **Data domain**
  Mix of proprietary and public data with a direct link to personally identifiable data
- **Acquisition of capabilities**
  Learn from provided data
- **System task**
  Forecasting – uses past and existing behaviours to predict future outcomes
- **Level of action autonomy**
  Medium (human on-the-loop) [in principle, but may be higher in practice?]



OECD.AI Policy Observatory

Next Steps: AI System Classification Framework

- Test framework's utility with human users
- Refine it based on input
- Populate it with more AI system examples
- Publish the updated version in early 2021

Draft classification made possible by…

Members:

57 experts participate in ONE-CAI.

Co-chairs:

- **Marko Grobelnik**, AI Researcher & Digital Champion, AI Lab, Slovenia Jozef Stefan Institute;
- **Dewey Murdick**, Director of Data Science, Center for Security and Emerging Technology (CSET), School of Foreign Service, Georgetown University; and
- **Jack Clark**, Policy Director, OpenAI.

Secretariat:

- The OECD team incl. Karine Perset, Luis Aranda, Nobu Nishigata
- Consultants including Tim Rudner, Doaa Abu-Elyounes, Peter Cihon

OECD.AI
Policy Observatory

Backup

OECD AI Systems Classification Framework

OECD.AI Policy Observatory

Objective:

- provide a structure to assess and classify AI systems according to their impact on public policy in areas covered by the OECD AI Principles.

Key points:

- The framework is simplified and user-friendly rather than exhaustive.
- The robustness and applicability of the present framework will be tested in late 2020 / early 2021 and adjustments made if needed.
- The 10 OECD AI Principles are used to structure the analysis of policy considerations associated with each dimension and sub-dimension

Values-based principles for all AI actors

1.1. Economic, social and environmental impact
1.2. Human rights including privacy, fairness
1.3. Transparency, explainability
1.4. Robustness, security, safety
1.5. Accountability

Recommendations to policy makers

2.1. Investment in research
2.2. Data, compute, technologies
2.3. Enabling policy and regulatory environment
2.4. Jobs, automation, skills
2.5. International cooperation