
GSC
22
MONTREUX, SWITZERLAND
IEC
ISO

# Transparent and trustworthy AI/Machine Learning

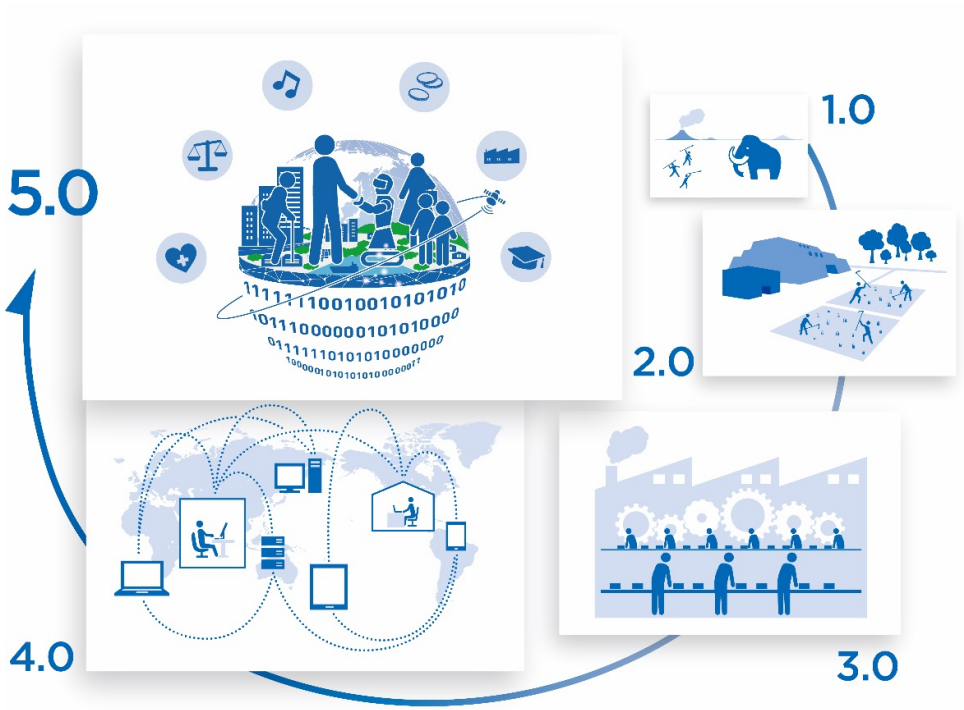
Yutaka Miyake,
TTC & KDDI Corporation, Japan

GSC-22, Montreux Switzerland, 26-27 March 2019

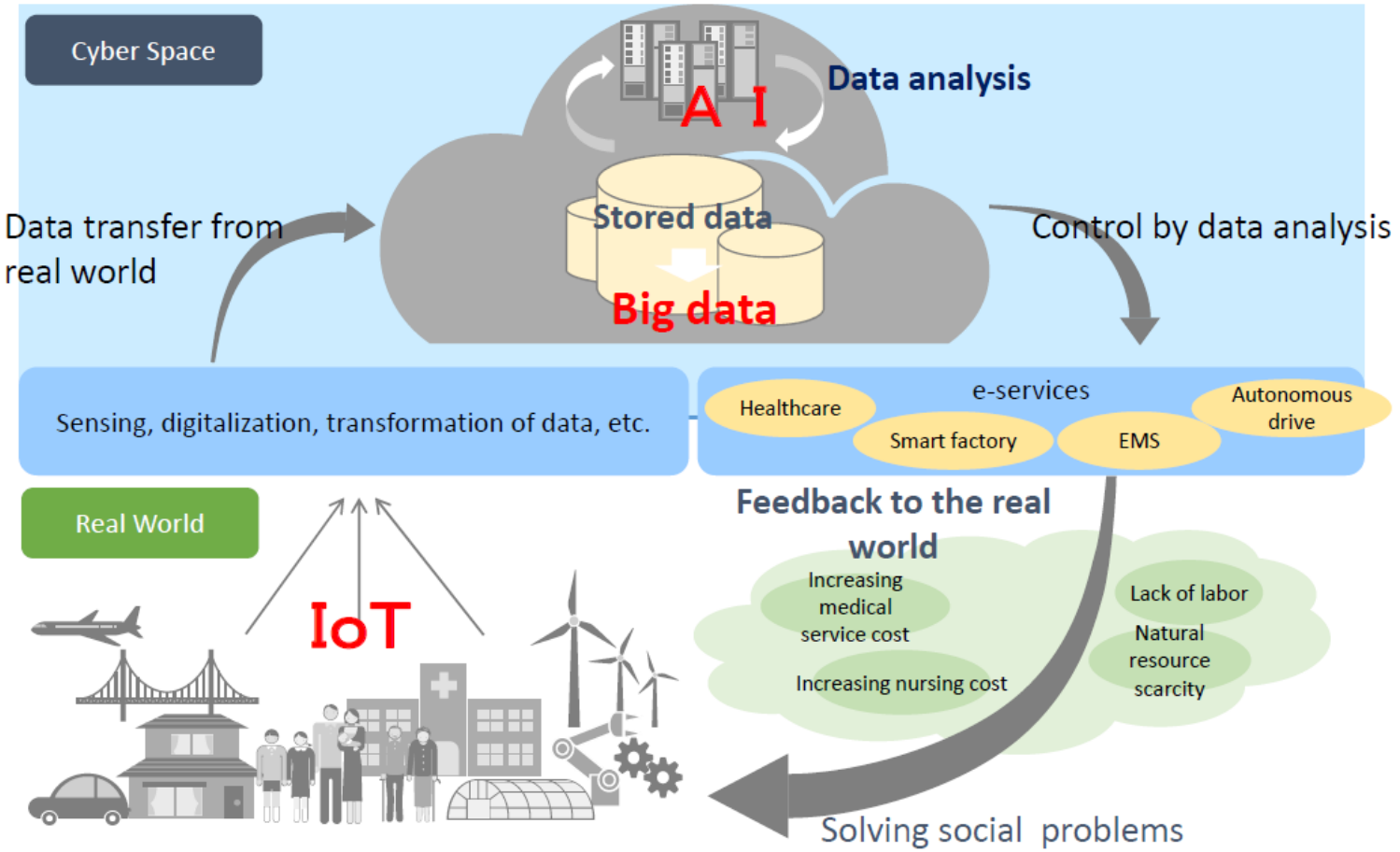
# AI/ML in society

# Super Smart Society “Society 5.0”

- The Japanese 5th Science and Technology Basic Plan, a comprehensive plan to promote science and technology in Japan over a five-year term (FY2016 to FY2020) adopted by the Cabinet in January 2016, aims to realize a future society “Society 5.0”.  The Society 5.0 makes the most of ICTs to facilitate human prosperity.

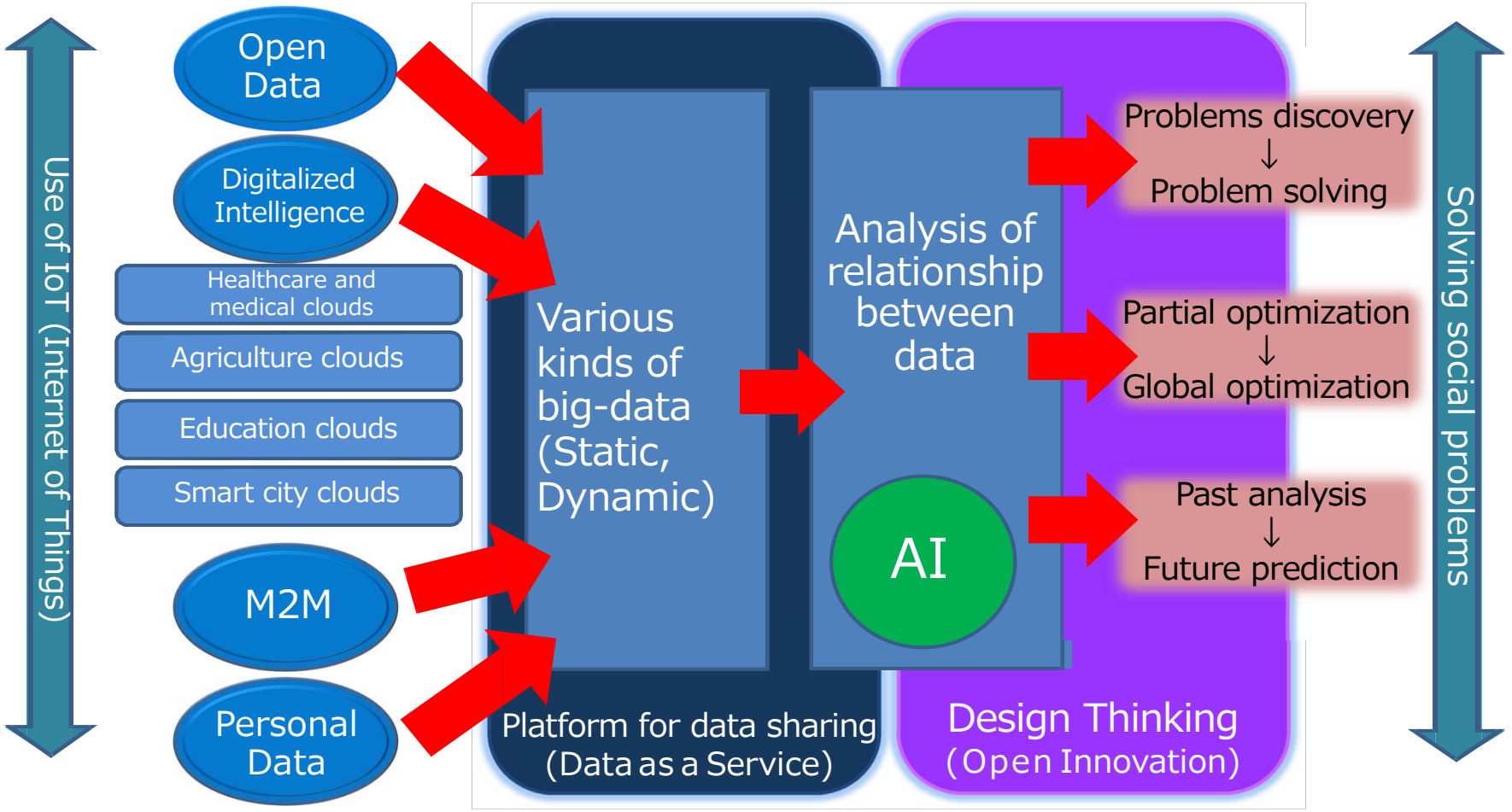


# Data circulation system for solving social problems




Ref: MIC (Ministry of Internal Affairs and Communications, Japan)

# Data circulation system for solving social problems



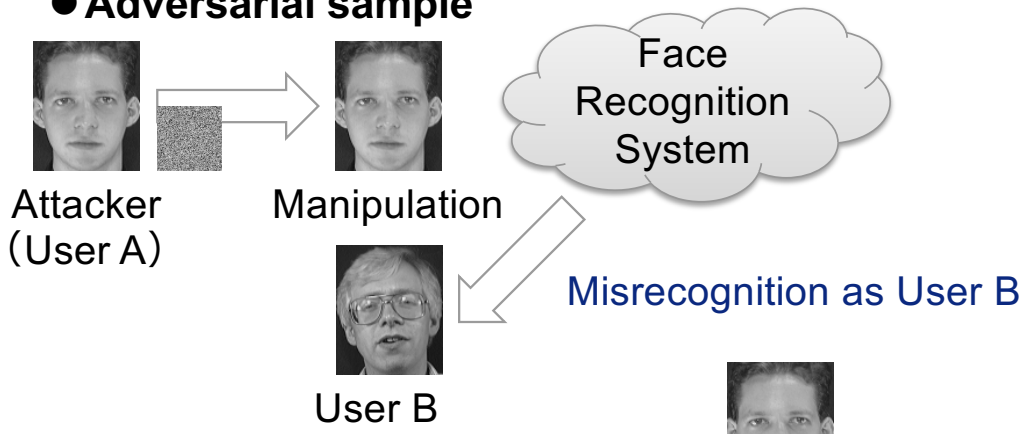

Ref: MIC (Ministry of Internal Affairs and Communications, Japan)

# Concerns on AI

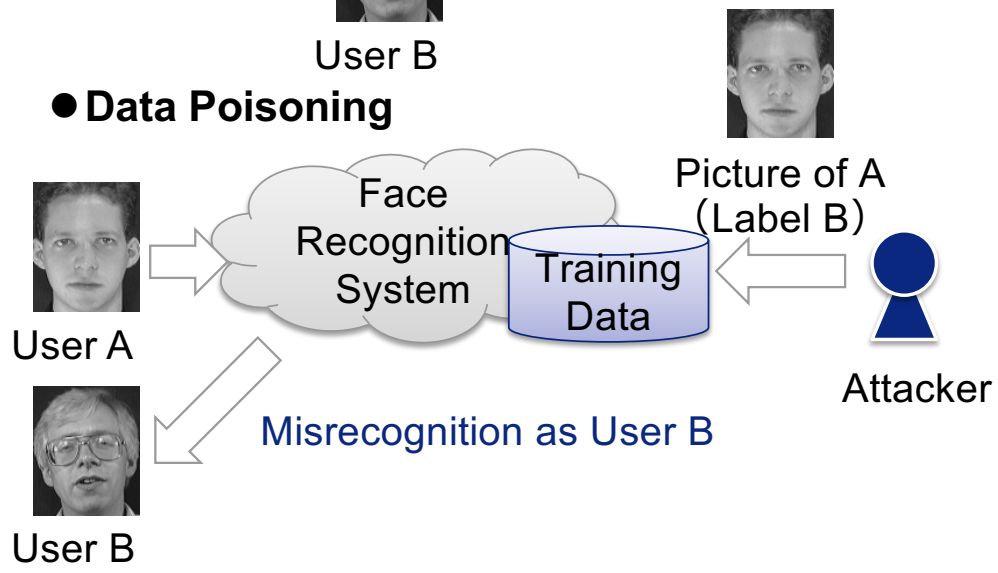
# Threat examples on AI

◆ **Security**

● **Adversarial sample**

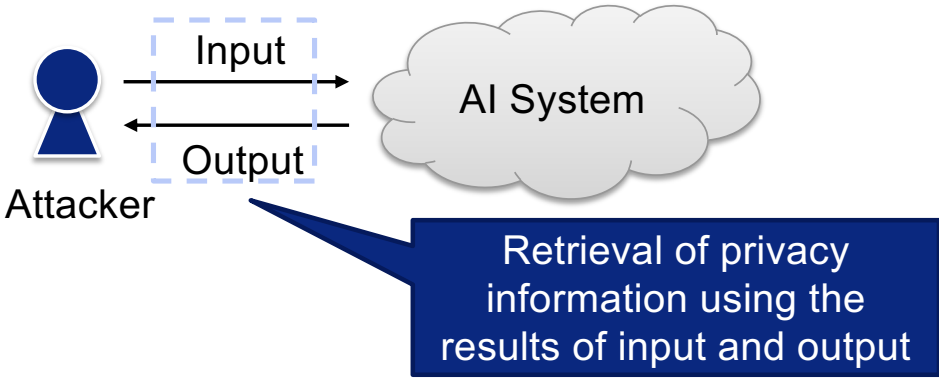



● **Data Poisoning**



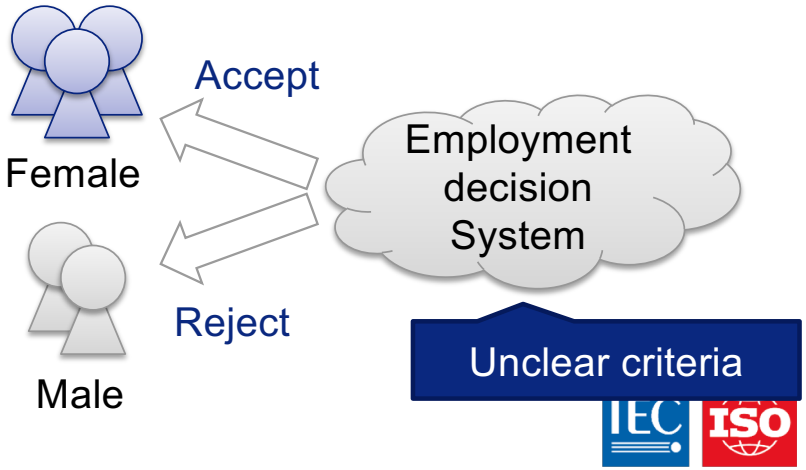

◆ **Privacy**

● **Model Inversion attack**

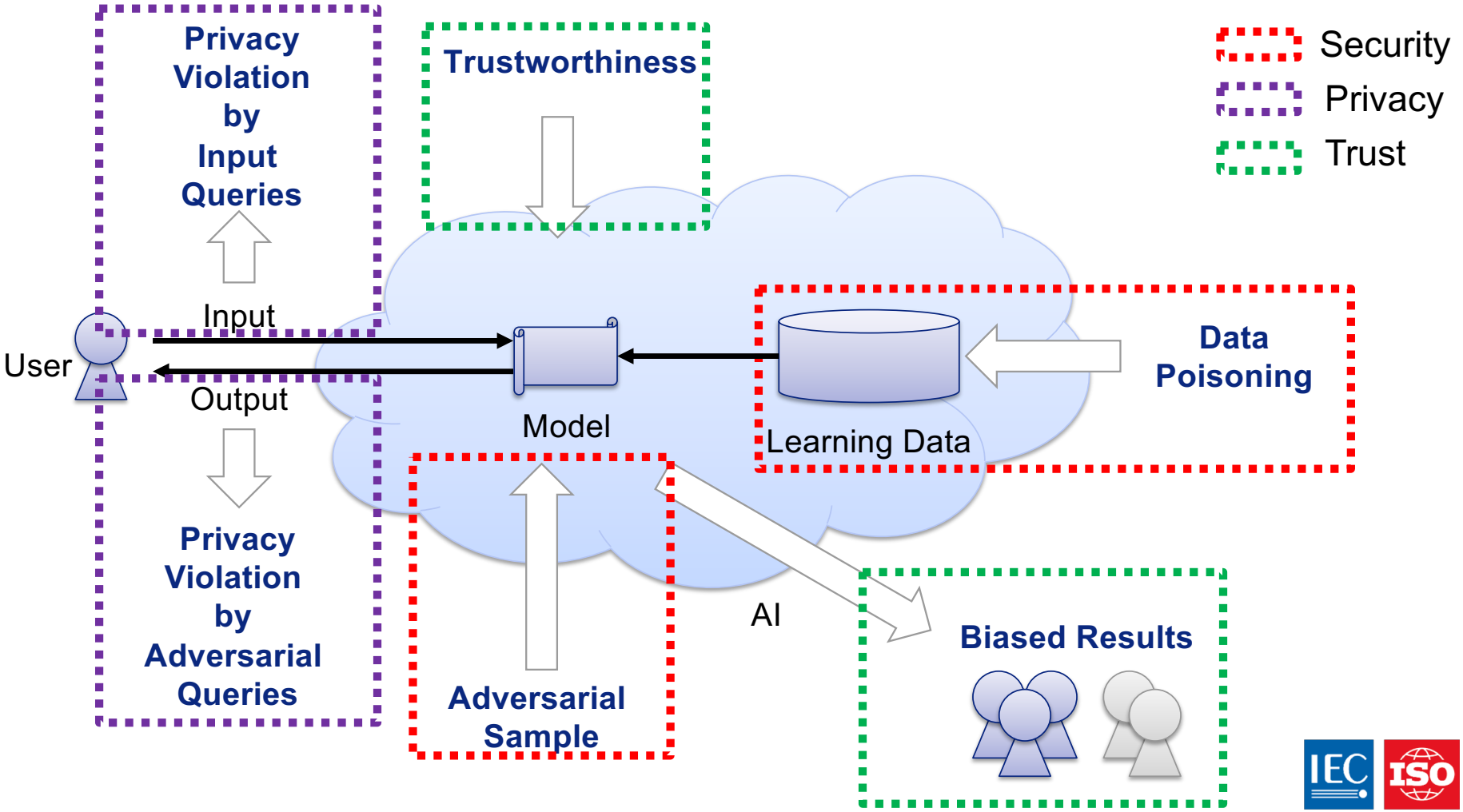



◆ **Trust**

● **Unfairness decision by AI**



# General concerns on AI

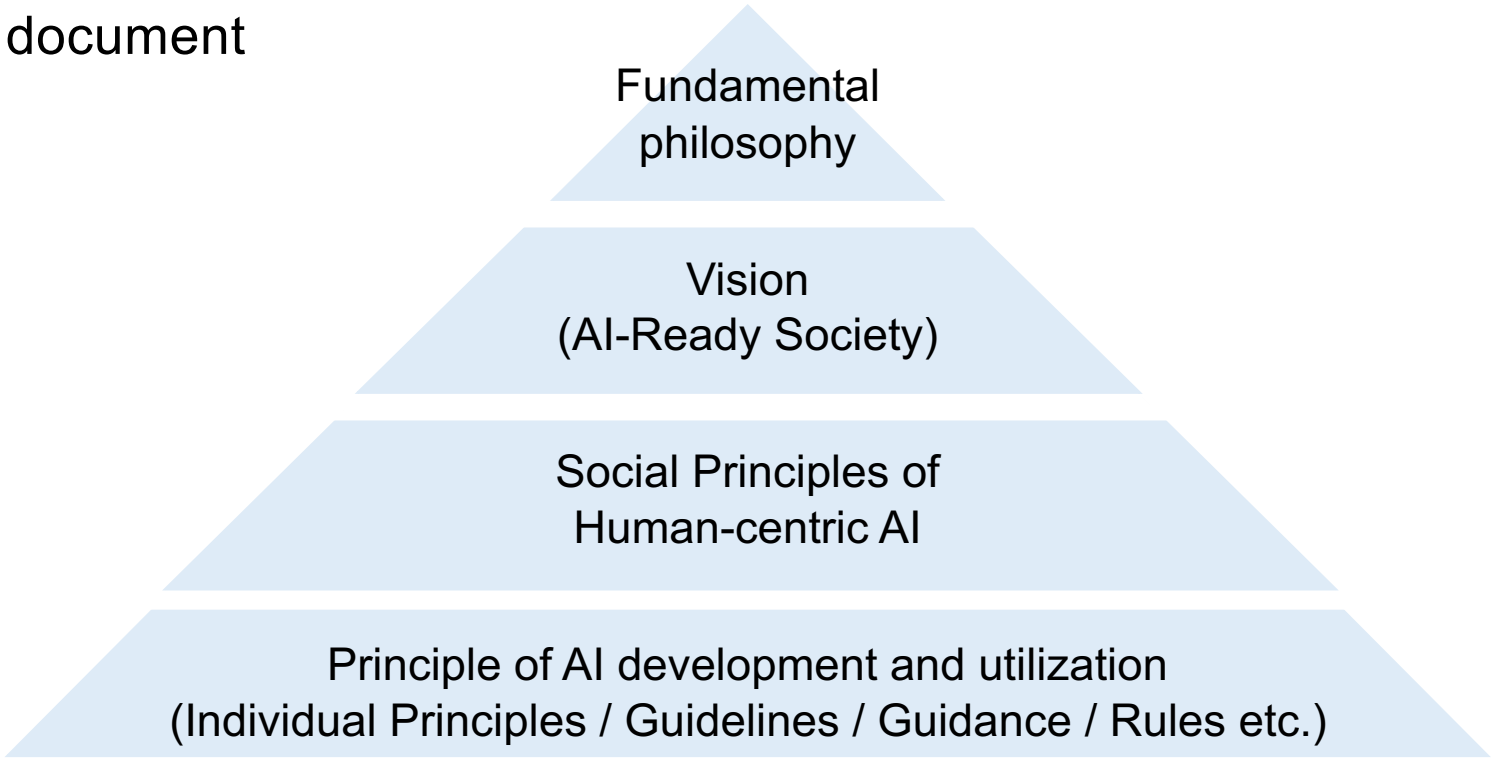


# Social Principles of Human-centric AI (Draft) – Dec. 27, 2018

(Japanese Government)

# Social Principle of Human-centric AI (Draft)

Structure of document




This principle is to be drawn up in March 2019 after soliciting domestic and international opinions.
(https://www8.cao.go.jp/cstp/stmain/20190115aigensoku.html)

# Social Principles of AI

1. Human-centric
   - Utilization of AI should not infringe upon fundamental human rights that are guaranteed by the Constitution and international norms.
2. Education
   - The following people should understand the AI, and user the AI properly.
     - Policy-makers and managers of the enterprises involved in AI
     - AI users
     - Developers of AI
3. Privacy
   - The personal data used in AI should be handled properly.
4. Security
   - There are new security risks for the user of AI. Society should always be aware of the balance of benefit and risks, and should work to improve social safety and sustainability as s whole.

# Social Principles of AI

5. Fair Competition
   - A fair competitive environment must be maintained to create new businesses and services, to keep economic growth sustainable, and to solve social issues.
6. Fairness, Accountability, and Transparency
   - Under the “AI-Ready society”, when using AI, fair and transparent decision-making and accountability for the results should be appropriately ensured, and trust in technology should be secured, in order that people using Ai will not be discriminated on the ground of the person’s background or treated unjustly in light of human dignity.
7. Innovation
   - To ensure the sound development of AI technology, it is necessary to establish and accessible platform in which data from all fields can be mutually utilized across borders with no monopolies, while ensuring privacy and security.

# Challenges on Transparency of AI

# Transparency

Explainable Artificial Intelligence [Wikipedia]

- Explainable AI (XAI), Interpretable AI, or Transparent AI refer to techniques in artificial intelligence (AI) which can be trusted and easily understood by humans. It contrasts with the concept of the "black box" in machine learning where even their designers cannot explain why the AI arrived at a specific decision. XAI can be used to implement a social right to explanation. Some claim that transparency rarely comes for free and that there are often tradeoffs between how "smart" an AI is and how transparent it is; these tradeoffs are expected to grow larger as AI systems increase in internal complexity. The technical challenge of explaining AI decisions is sometimes known as the interpretability problem. Another consideration is info-besity (overload of information), thus, full transparency may not be always possible or even required. The amount of information presented should vary based on the stakeholder interacting with the intelligent system.

# Interpretability

- For works that describe machine learning models as black boxes, **transparency and interpretability** are closely related, if not the same concept.

- Common approach proposed to address the opacity of models is through improving that interpretability
- **Post hoc interpretability**
  - Aims to explain the resulting prediction of black box models
- **Interpretable models**
  - Introduce interpretability in the model itself

# Research about Post hoc interpretability

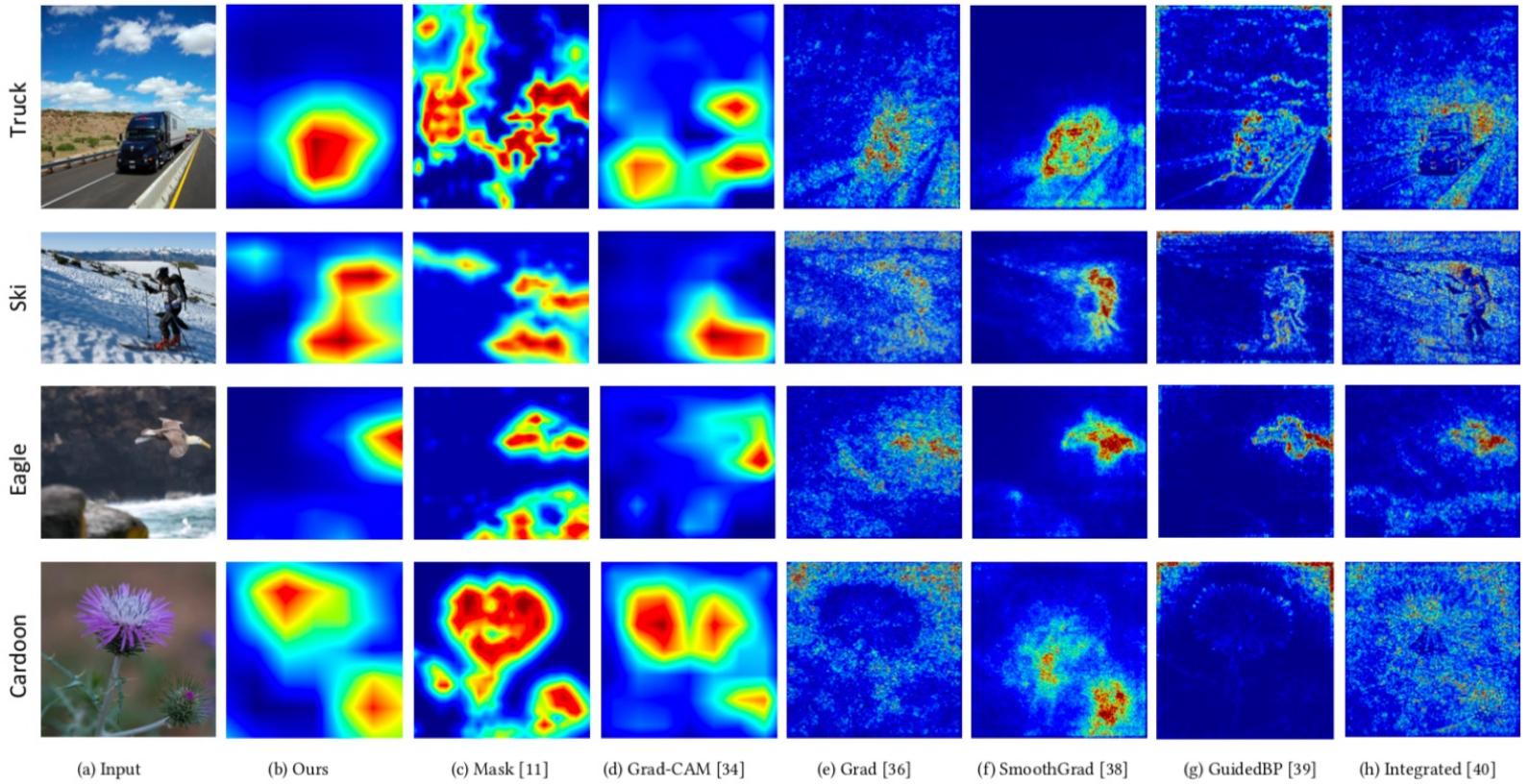
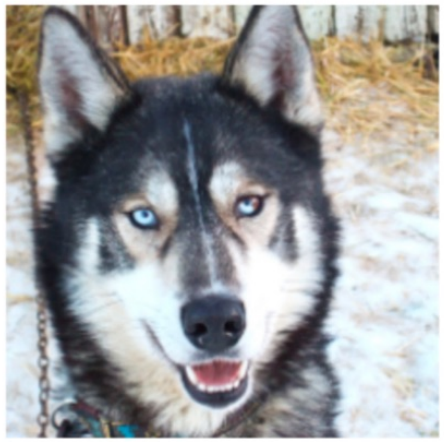



**Figure 2: Visualization saliency maps comparing with 6 state-of-the-art methods.**

Du, M., Liu, N., Song, Q., & Hu, X. (2018). Towards Explanation of DNN-Based Prediction with Guided Feature Inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1358–1367

# Interpretability and Trustworthiness

- If the results of a model can be interpreted, that provides information that can help decide on the trustworthiness of the model



(a) Husky classified as wolf

(b) Explanation

**Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.**

| | Before | After |
|---|---|---|
| Trusted the bad model | 10 out of 27 | 3 out of 27 |
| Snow as a potential feature | 12 out of 27 | 25 out of 27 |

**Table 2: "Husky vs Wolf" experiment results.**

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

# Transparency Risks

- **Security**
  - Transparency may make **models vulnerable to attack** by increasing the understanding regarding how the results are obtained (Papernot, McDaniel, Sinha, & Wellman, 2018).
- **Privacy**
  - Transparency could reveal private information. Rule disclosure may be prohibited by law if it involves private information. (Kroll et al. 2016; Ananny & Crawford 2018)
  - Should consider what, how and to whom the information is revealed.
- **Intellectual property** (Papernot, McDaniel, Sinha, & Wellman, 2018).

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and Privacy in Machine Learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS P)* (pp. 399–414)
Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989.
Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. U. Pa. L. Rev., 165, 633.

# Standardization on AI

# What will be required for ‘AI Security’ Standardization?

- **Background**
  - There are several important concerns on AI.
  - Governments recommend AI users to follow their principles of AI system usage.
  - The users would like to know the criteria to satisfy the principles.
- **What items will be required?**
  - How to align requirements from government principles on AI.
  - Guidelines
    - Guideline for using AI system
    - Reporting method of AI system
  - Technical Specifications
    - Internal evaluation of training data
    - Investigation of output data for standardized training data
    - Detection of unbalanced output
    - Trust framework for using AI