Current Status and Initiative of AI in Japan

2024-10-17

Director General of Digital Infrastructure Center
Deputy Executive Director and Secretary General of AI Safety Institute
Information-technology Promotion Agency, Japan

Kenji Hiramoto

IPA IT Promotion Agency Japan


IPA is not beer!


IPA is
the government-funded agency
for the digital technologies.

Overview of our activities

IPA Japan

Anyone can start new businesses easily.

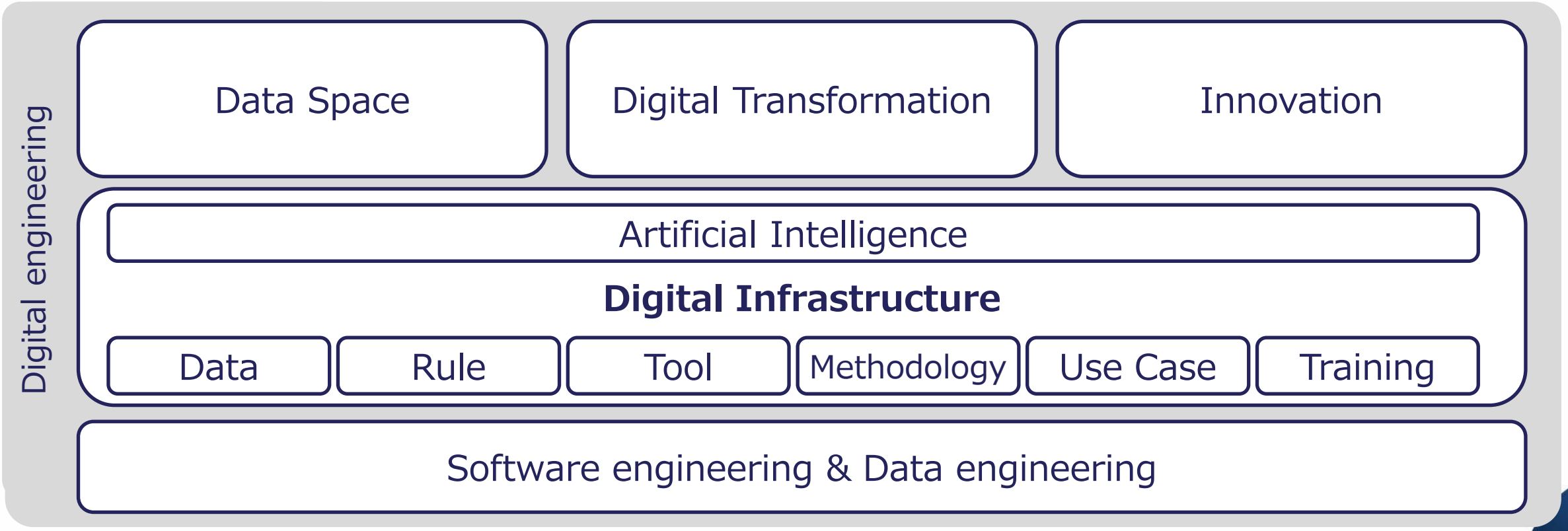
Anyone can transform their businesses.

Anyone can realize their ideas.

Design data spaces

Digitalize & revolute businesses

Incubate talents and tech-ideas

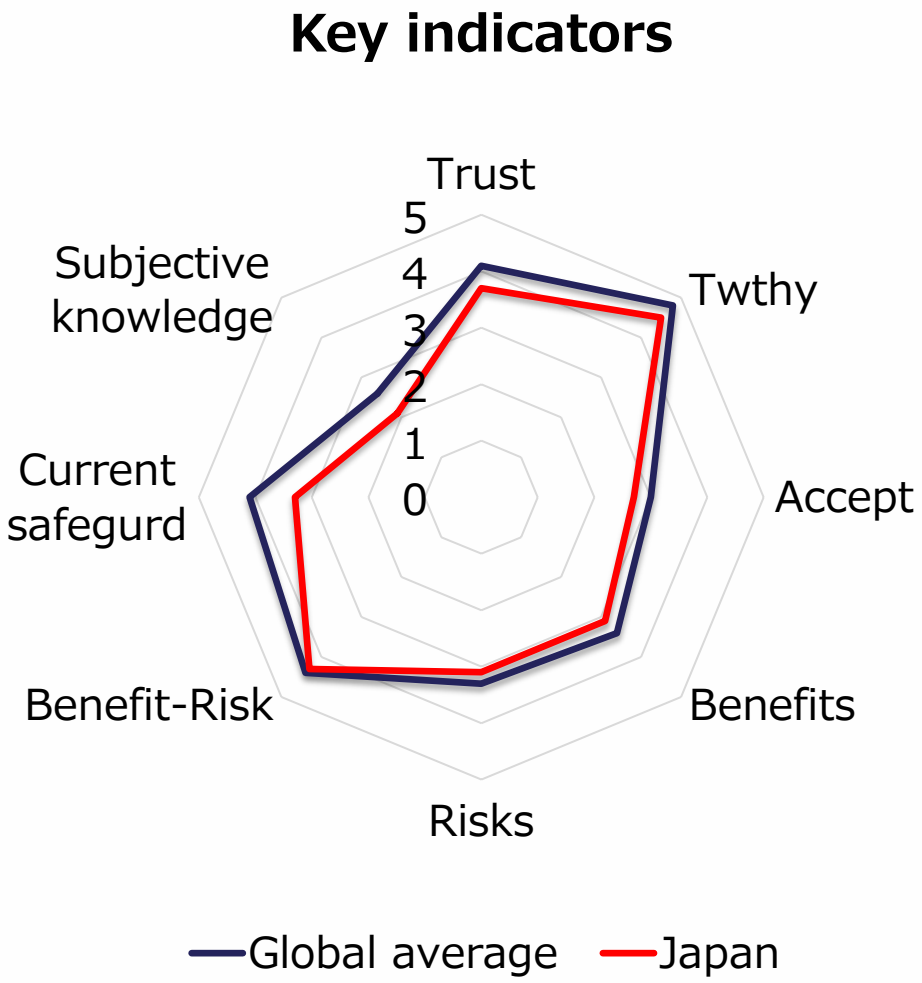


Human Resource

Security

AI in our life and business

Citizen's Perspective

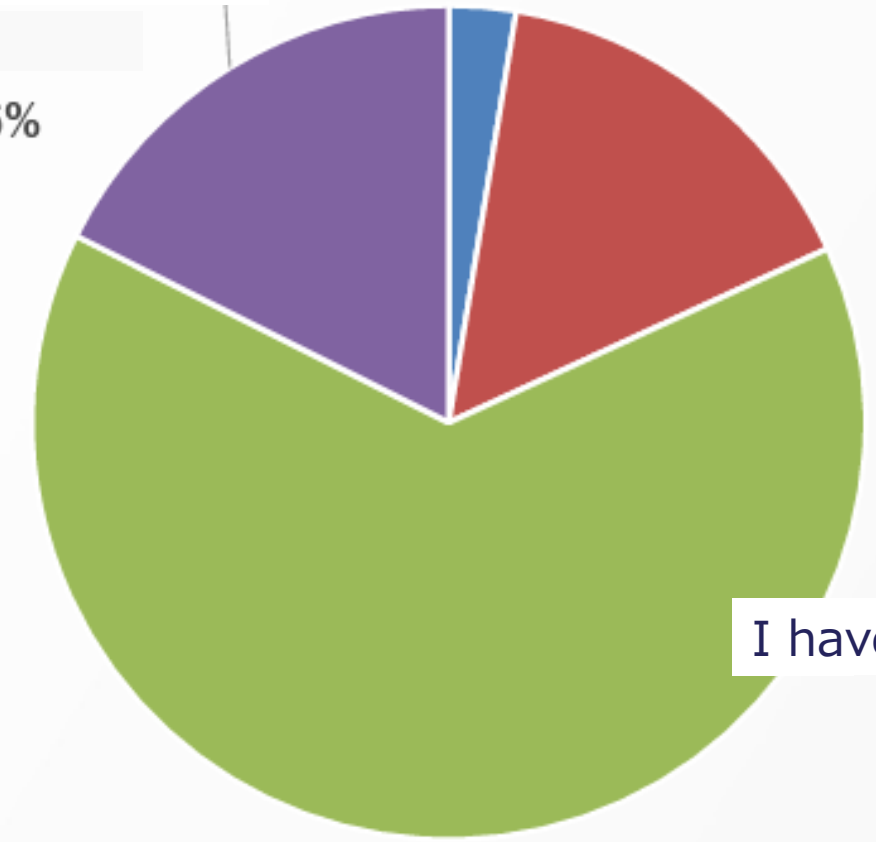


- The key indicators show us the similar trend.
- Most of citizens recognize the benefits and risks of AI.
  - AI provides benefit :75%
  - AI is reliable :23%
  - AI has risks :68%
  - I don’t have enough knowledge of AI :75%
  - I’m using applications include AI :56%

Notes:
Trust = Trust in AI systems,
Twthy = Perceived trustworthiness of AI systems,
Accept = Acceptance of AI systems,
Benefits = Perceived benefits of AI systems,
Risks = Perceived risks of AI systems,
Benefit-Risk = Perception that benefits of AI systems outweigh the risks,
Current Safeguards = Perceived adequacy of current laws and regulations governing AI,
Subjective Knowledge = Self-reported knowledge of AI.

Source : Trust in Artificial Intelligence :A global study 2023, KPMG
https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/trust-in-ai-global-insights-2023.pdf

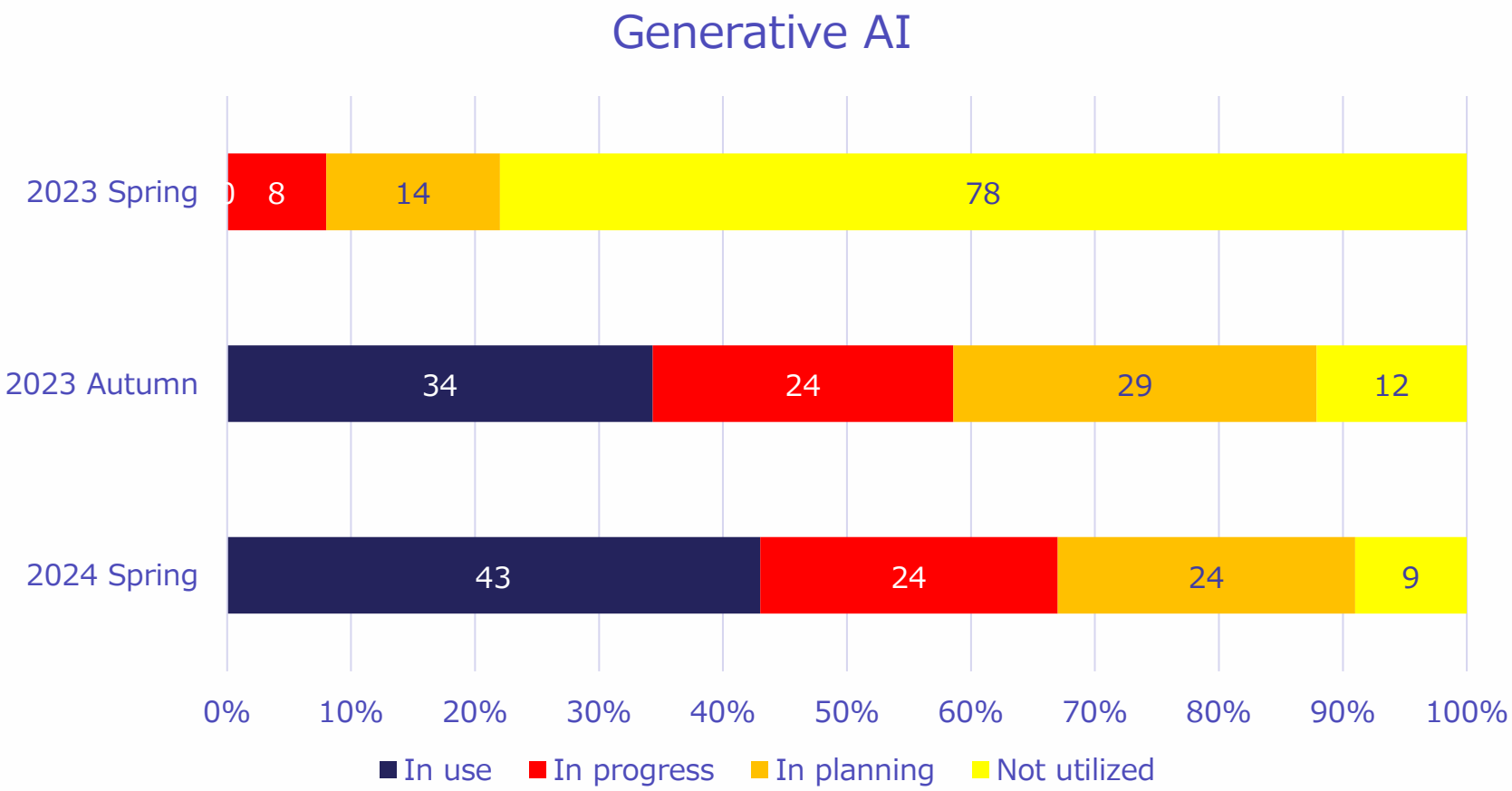
Generative AI in daily life



デジタル社会における消費者意識調査2024, JIPDEC
https://www.jipdec.or.jp/news/pressrelease/m0p0h60000005qig-att/20240418_01.pdf

Generative AI takes off

- The use of generative AI in Japan has been on the rise rapidly.



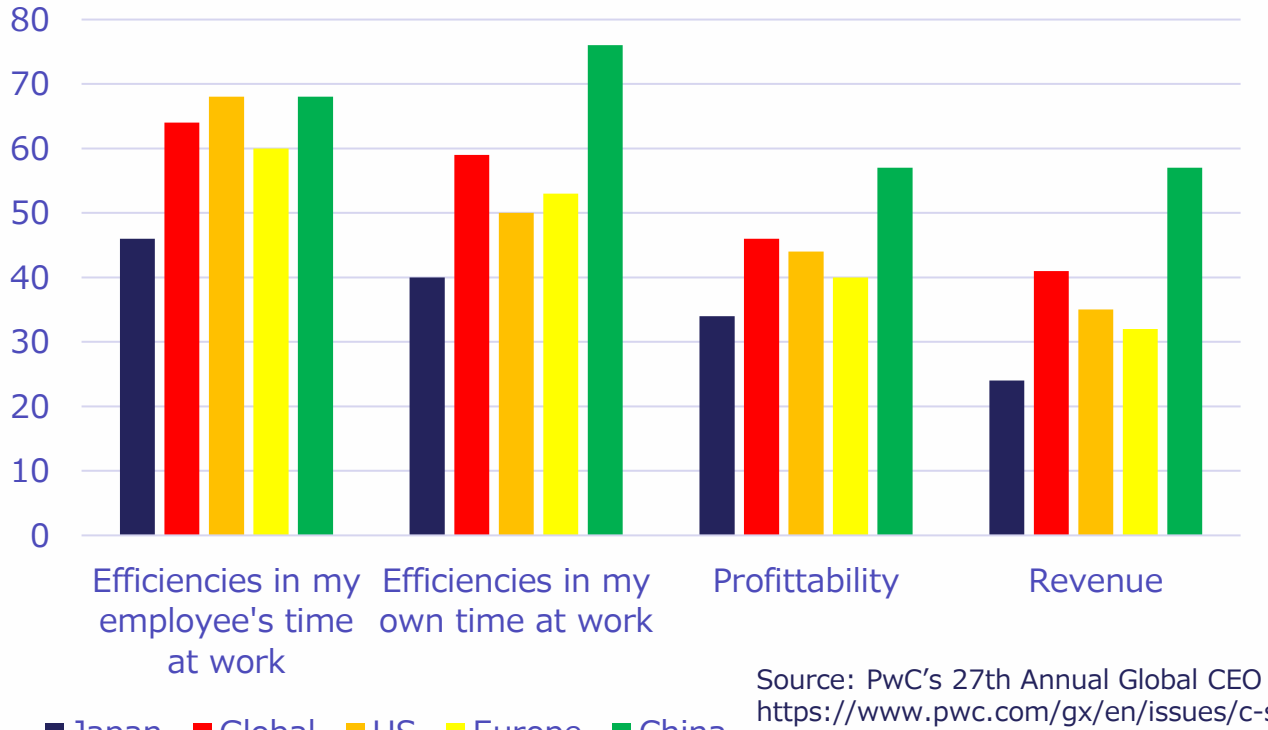
Source: generative AI survey 2024 Spring(PWC)
https://www.pwc.com/jp/ja/knowledge/thoughtleadership/generative-ai-survey2024.html

Business(CEO) Perspectives

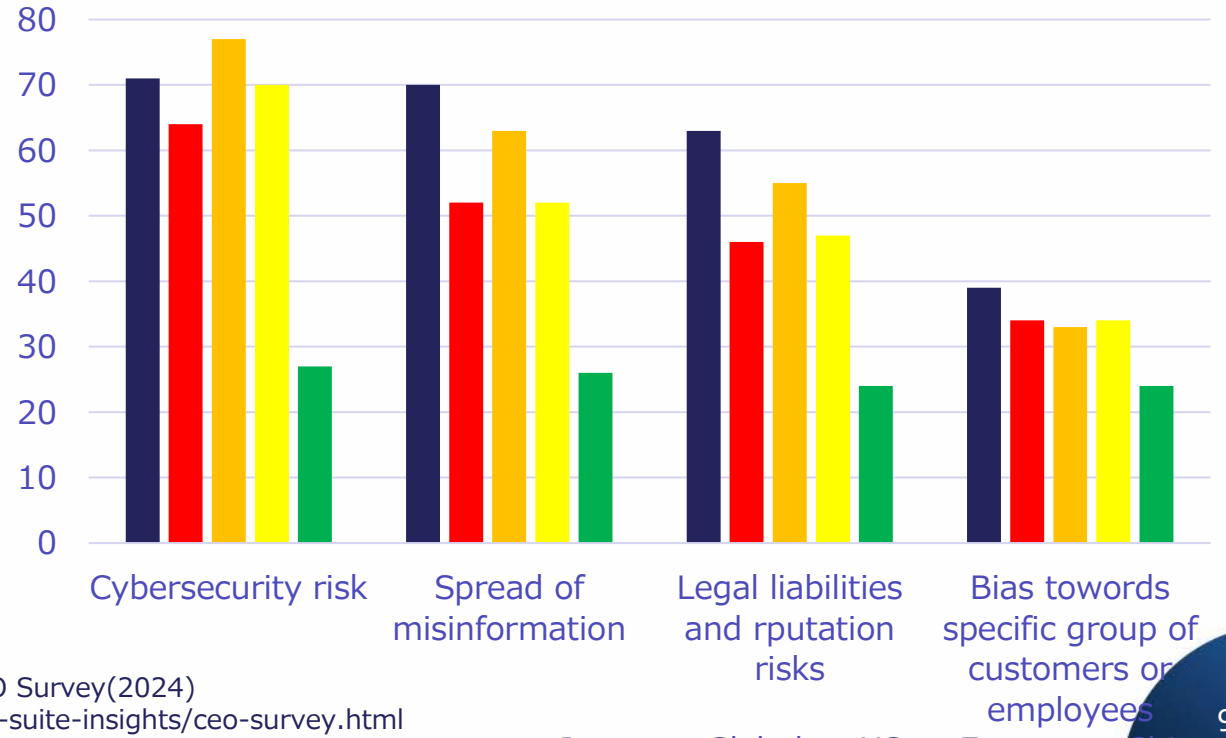
IPA Japan

- Japanese CEOs have yet to realize the benefits of generative AI. On the other hand, they feel the risks.

To what extent will generative AI increase the following in your company



To what extent do you agree that generative AI is likely to increase the following in your company in the next 12 months?



Source: PwC's 27th Annual Global CEO Survey(2024)
https://www.pwc.com/gx/en/issues/c-suite-insights/ceo-survey.html
https://www.pwc.com/jp/ja/knowledge/thoughtleadership/ceo-survey.html

Barriers of implementing generative AI

- Lack of human resources and experience is the biggest challenge. Another problem pointed out is the lack of basic social data.

Source: generative AI survey 2023autumn(PWC)
https://www.pwc.com/jp/ja/knowledge/thoughtleadership/generative-ai-survey2023_autumn.html

AI Strategy and policy

Innovation Strategy 2024

IPA Japan

- AI safety is an essential requirement for an environment that encourages innovation.

① AI innovation and the acceleration of innovation through AI
- Strengthening R&D capabilities (including data supply)
- Acceleration of the use of AI
- Enhancing AI infrastructure
- Human resource development and recruitment

② Realize the AI Safety
- Governance and rules
- AI safety
- Prevention of mis/dis information
- Intellectual property rights

③ International cooperation/collaboration

Overview of the AI Safety Institute (AISI)

AISI : AI Safety Institute

AISI Japan AI Safety Institute

- **Objectives**
  - **AISI supports public and private sector efforts.**
    - The public and private sectors need to work together to ensure that all parties involved in the development and use of AI are properly aware of the risks of AI. Governance also needs to be ensured throughout the lifecycle. Then, the safe and secure use of AI will be promoted.
    - Need to promote innovation and mitigate risks in the lifecycle, in those efforts.
- **Principles**
  - **AISI activities to be harmonized with related domestic and international organizations.**
    - Response to rapidly and globally advancing technologies.

Role and Scope of AISI

AISI Japan AI Safety Institute

- **Role**
  - **AISI supports the government** by conducting surveys on AI safety, examining evaluation methods, and creating standards.
  - **As a hub for AI safety in Japan**, AISI will consolidate the latest information in industry and academia, and promote collaboration among related companies and organizations.
  - **Collaborate with AI safety-related organizations**.
    - AISI is not an R&D organization.
- **Scope**
  - **Set the scope flexibly** in the following AI related issues, while considering **global trends**.
    - **Social Impact**
    - **governance**
    - **AI System**
    - **contents**
    - **data**

Activities

AISI Japan AI Safety Institute

1) Consideration of surveys and standards for AI safety assessment

   (i) Surveys on standards of AI safety, checking tools, anti-disinformation technology, AI and cybersecurity

   (ii) Consideration of standards and guidance related to AI safety

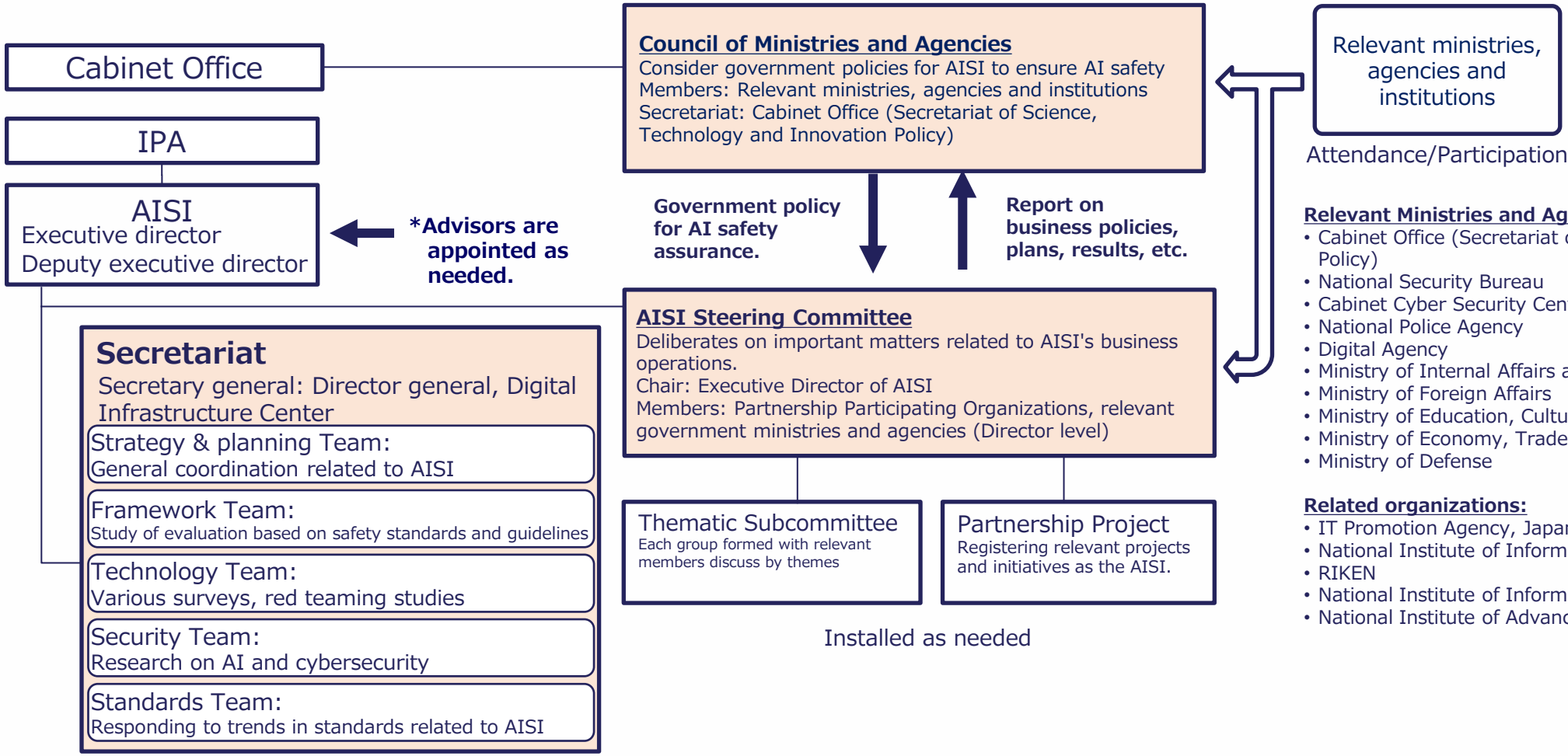
   (iii) Consideration of a testbed environment for AI related to the above

2) Consideration of implementation methods for AI safety assessment

3) International collaboration with related organizations in other countries (such as the AI Safety Institute in the U.K. and the U.S.)

AISI Structures

- "Council of Ministries and Agencies", set up in Cabinet Office, deliberates on the important matters of AIS1.
- The "AISI Steering Committee" within AISI reports to the Council (to be held once a month). Under the Steering Committee, "thematic subcommittees" and "partnership projects" will be installed as necessary.
- As the secretariat of AISI, five teams were formed within the IPA Digital Infrastructure Center.



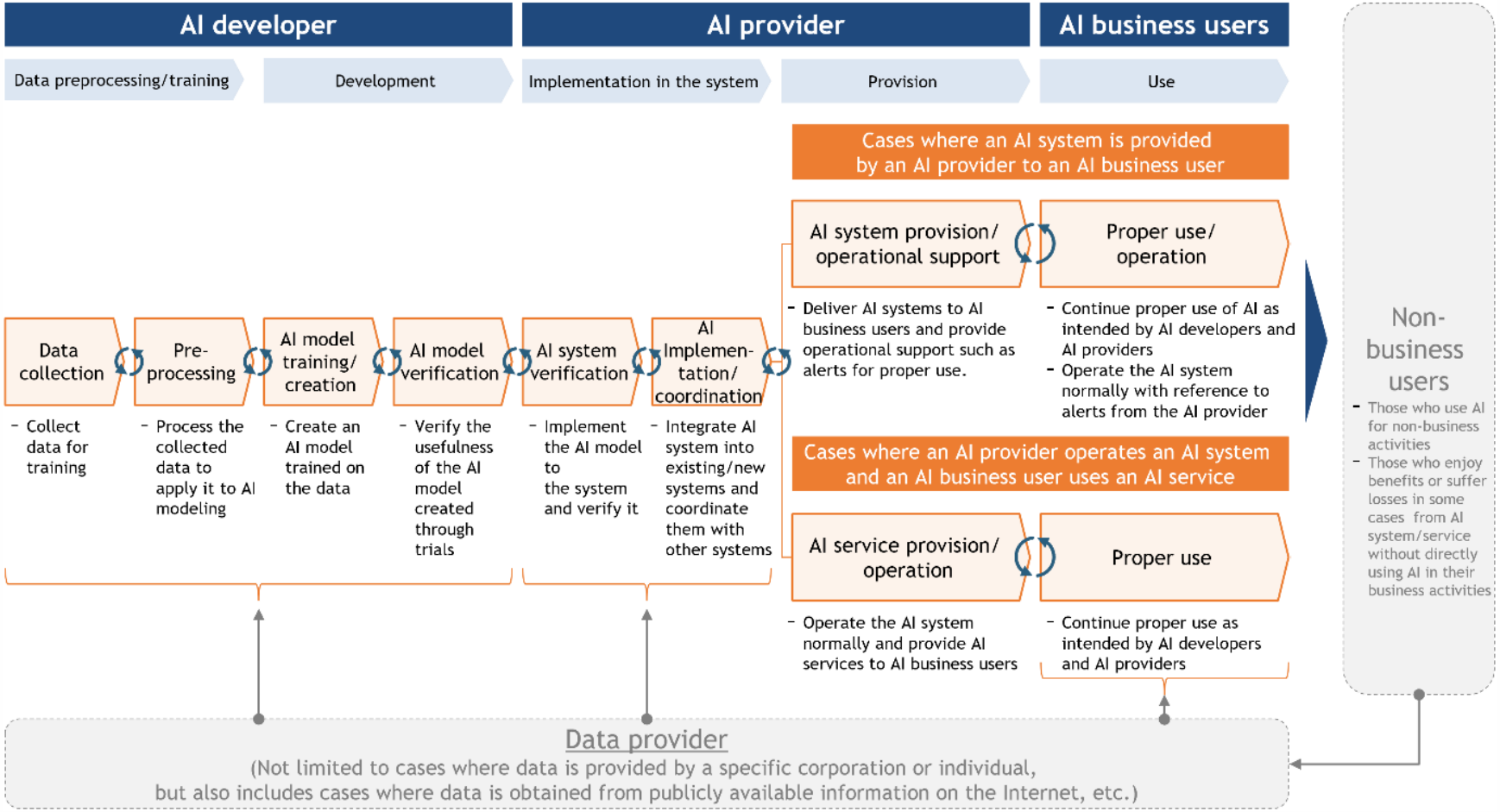
Relevant Ministries and Agencies:
- Cabinet Office (Secretariat of Science, Technology and Innovation Policy)
- National Security Bureau
- Cabinet Cyber Security Center
- National Police Agency
- Digital Agency
- Ministry of Internal Affairs and Communications
- Ministry of Foreign Affairs
- Ministry of Education, Culture, Sports, Science and Technology
- Ministry of Economy, Trade and Industry
- Ministry of Defense

Related organizations:
- IT Promotion Agency, Japan (IPA)
- National Institute of Information and Communications Technology
- RIKEN
- National Institute of Informatics
- National Institute of Advanced Industrial Science and Technology

AI Guidelines for Business

AISI Japan AI Safety Institute

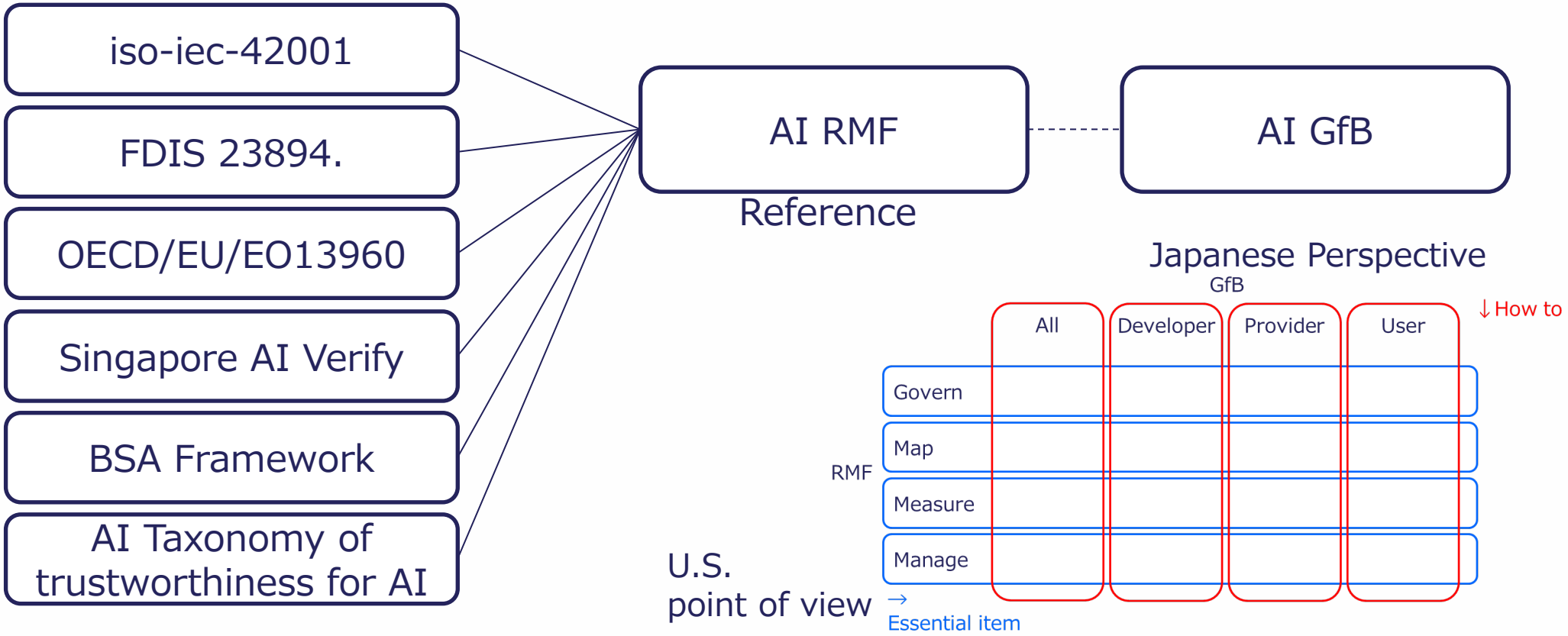
- Clarify what each stakeholder should address in the flow of AI utilization



Japan-U.S. Crosswalk

AISI Japan AI Safety Institute

- Confirmation of the interrelationship between the U.S. NIST AI Risk Management Framework (RMF) and the Japanese AI Guidelines for Business (GfB)



Other guidelines

IPA Japan

Guide for Evaluation Points Regarding AI Safety

Key Elements of AI Safety	Evaluation Perspectives on AI Safety: Control of Toxic Output	Prevention of Misinformation, Disinformation and Manipulation	Fairness and Inclusion	Addressing to High-risk Use and Unintended Use	Privacy Protection	Ensuring Security	Explainability	Robustness	Data Quality	Verifiability
Human-centric	●	●	●	●						
Safety	●	●		●				●	●	
Fairness	●		●						●	
Privacy protection					●					
Ensuring security						●				
Transparency		●	●				●	●	●	●

※ Various studies on AI Safety evaluations are ongoing domestically and internationally across diverse fields in industry, government, and academia, and the situation is constantly changing. Therefore, this document presents the evaluation perspectives that are considered to be particularly important. The perspectives described in this document are not exhaustive and are expected to be updated in the future.

Types of Red Teaming

➢ Red teaming can be categorized as follows.

Category of red teaming tests based on prior knowledge of the attack planner/conductor

- **Black-box Test**
  (The attack planner/conductor does not have any prior knowledge of the system, such as its internal structure.)
- **White-box Test**
  (The attack planner/conductor has sufficient knowledge of the system, such as its internal structure.)
- **Gray-Box Test**
  (The attack planner/conductor has partial knowledge of the system, such as its internal structure.)

Category of the environment in which red teaming is conducted

- **Production Environment**
  (Production environment where AI systems are actually put into practice)
- **Staging Environment**
  (Environment for testing and checking for defects in conditions similar to those of the actual production environment)
- **Development Environment**
  (Environment for developing AI systems)

Category of how attack signatures are attempted

- **Red Teaming with Automated Tools**
- **Manual Red Teaming**
- **Red Teaming with AI Agents**

Typical attack methods on LLM systems

➢ Examples of typical attack methods against LLM systems. They should be considered in Red Teaming.

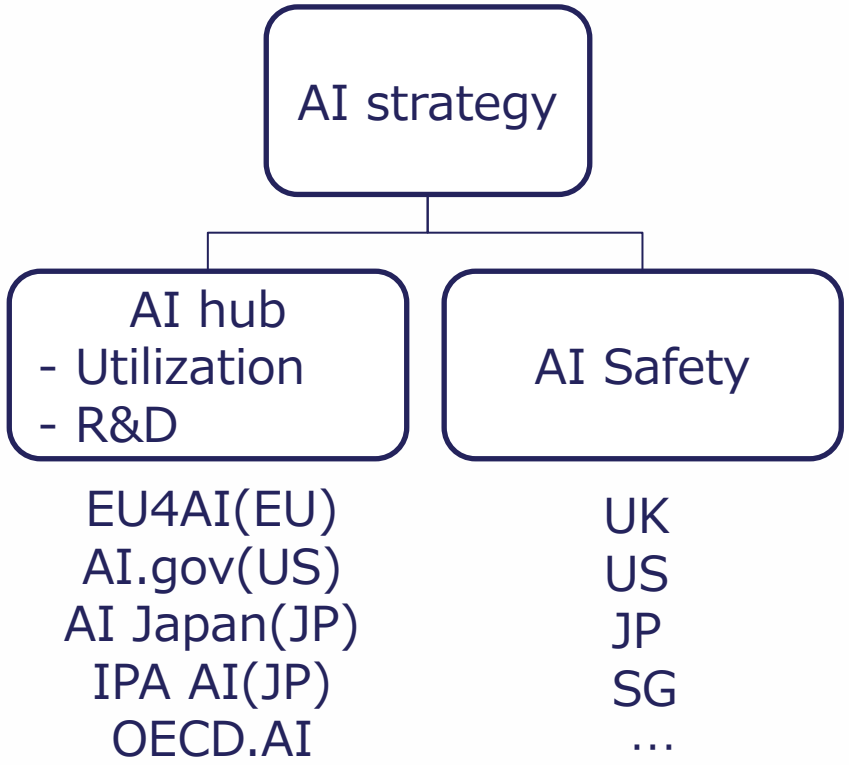
- **Direct Prompt Injection**
  Attacker directly injects malicious prompts into the AI system
- **Indirect Prompt Injection**
  Attacker indirectly injects malicious prompts into the AI system
- **Prompt Leaking**
  Attacker extracts the designated system prompt
- **Poisoning Attacks**
  Attacker infiltrates manipulated data or model into data or model during training
- **Evasion Attacks**
  Malicious modification of inputs to the AI system to cause unintended behavior
- **Model Extraction Attack**
  An attack to create a model with the same performance as the target system's model by analyzing its inputs and outputs
- **Membership Inference Attacks**
  An attack that identifies whether certain data is included in the training data by analyzing system's inputs and outputs
- **Model Inversion Attacks**
  An attack that recovers information contained in training data by analyzing inputs and outputs

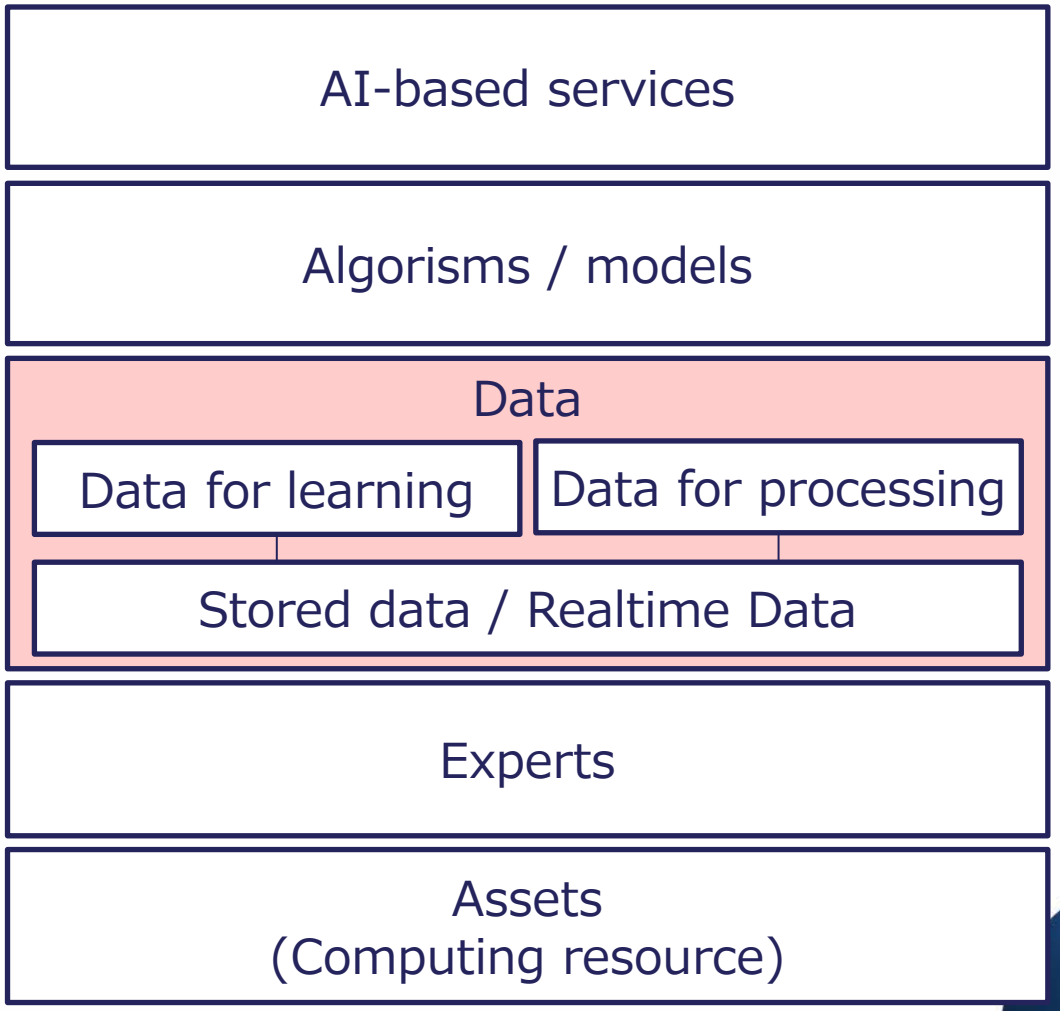
Red Teaming Methodology Guide for AI Safety

Data Infrastructure for AI

Data is essential for AI

- Every country and organizations strongly promote AI.
- One of the essential factor is data.





- AI Safety is focused topics.

There are many requests from AI team

IPA Japan

Issues

- High-quality data
- Trustful source
- Bias-less data
- Log management
- Traceable learning data
- Intellectual property
- Prevention of mis/dis information
- Ethics

Platform

- Data catalog
- Data market

Findability

- Ontology
- Data cleansing tool
- Data verification tool

Quality

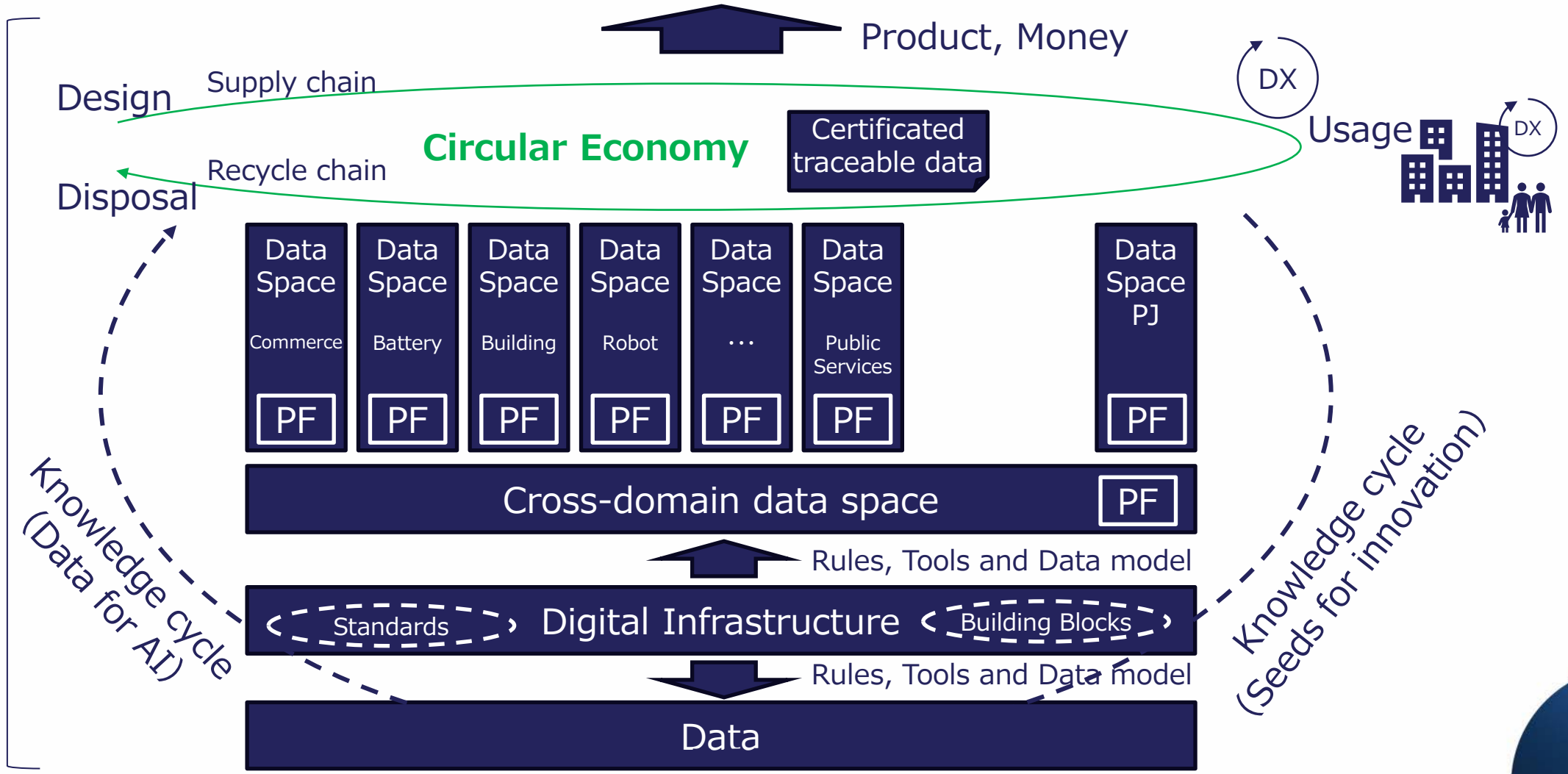
- Trust-related tool
- Log management tool
- Trace management tool

Trust

Data Ecosystem

IPA Japan

Goal: Competitive economy and well-being life
(Anyone launch and use digital services, anytime)

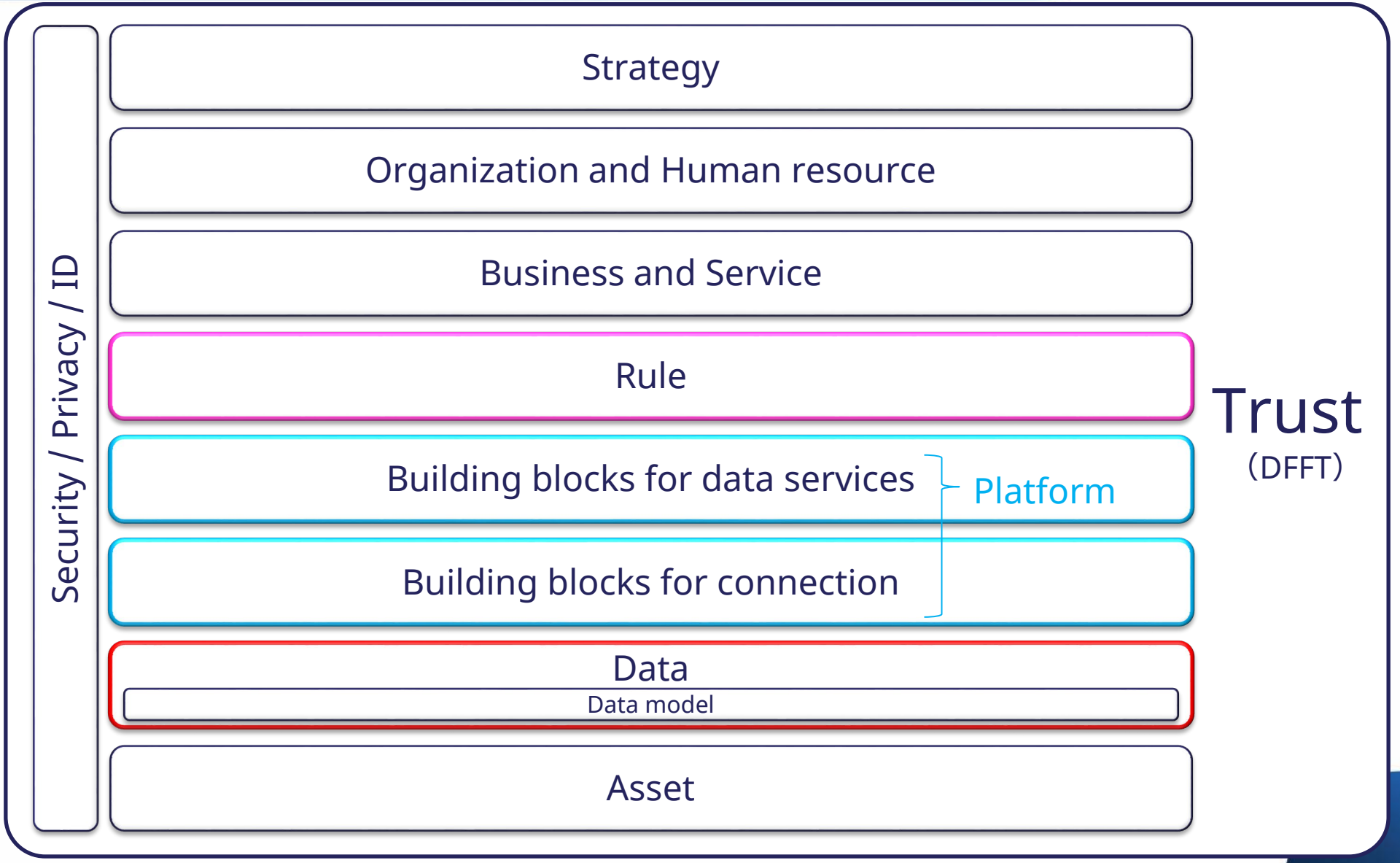


GIF(Government Interoperability Framework)

IPA Japan

- GIF is a part of our National Data Strategy.
- We define the GIF architecture.
- It especially focus on the Rule, Platform and Data layer.

GIF provide rules, tools, data models and technical guidelines.



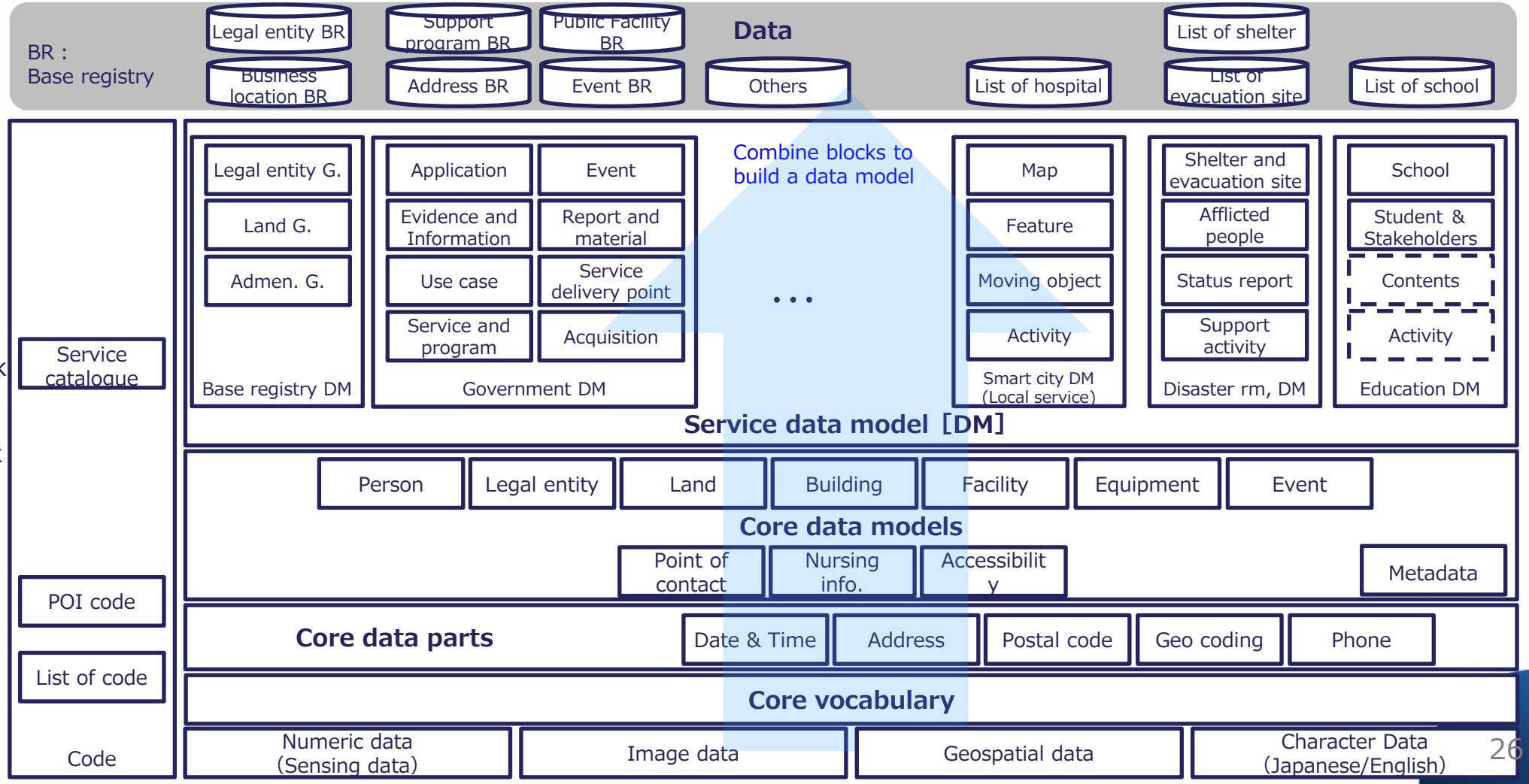
Guidebooks and Data models

- GIF Data Models and guidebooks support to make structured and high-quality data.

Guidebooks

GIF Guidebook

Character data guidebook
Master data design guidebook
Code design guidebook
API guidebook
Data management guidebook
Data specialist guidebook
Architecture guidebook
Data quality guidebook
Metadata guidebook



Overview of our activities

IPA Japan

Anyone can start new businesses easily.

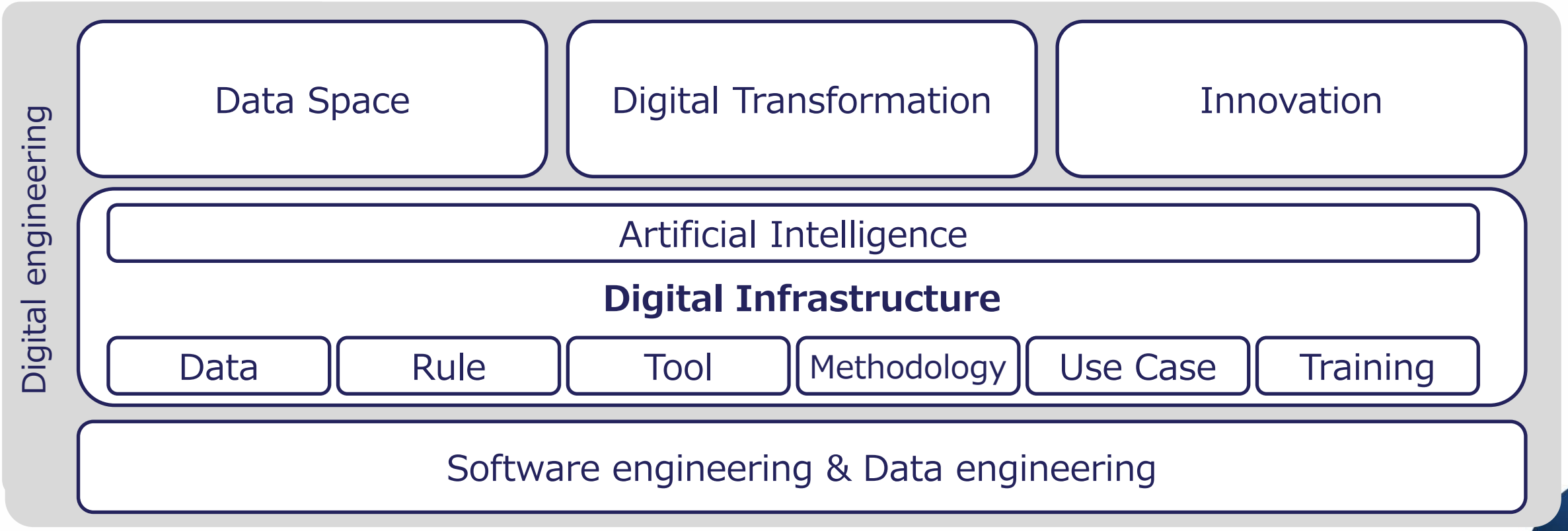
Anyone can transform their businesses.

Anyone can realize their ideas.

Design data spaces

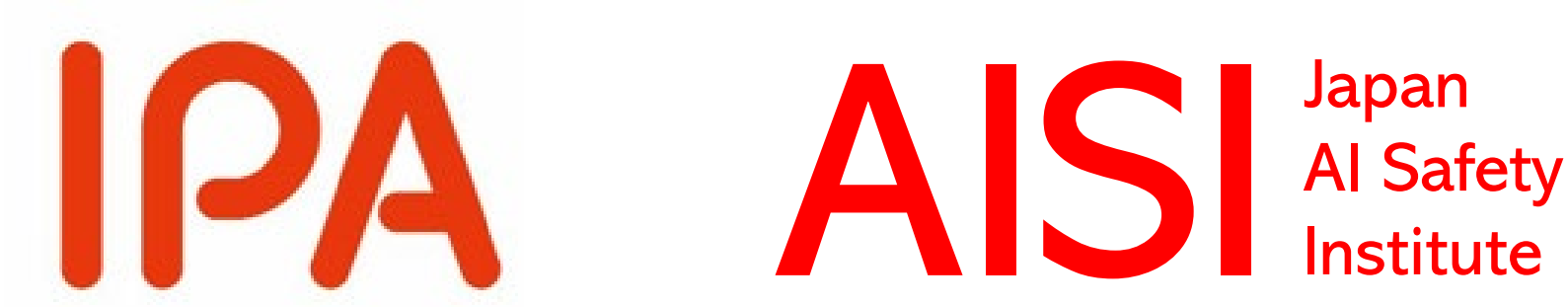
Digitalize & revolute businesses

Incubate talents and tech-ideas



Human Resource

Security


IPA
AISI
Japan
AI Safety
Institute